UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO DE ENGENHARIA DE SOFTWARE

NATHIELY LAIANE MORAES MACEDO

ABORDAGENS COMPUTACIONAIS PARA BIG DATA APLICADAS AO MERCADO DE AÇÕES: UM MAPEAMENTO SISTEMÁTICO

TRABALHO DE CONCLUSÃO DE CURSO

CORNÉLIO PROCÓPIO
2021

NATHIELY LAIANE MORAES MACEDO

ABORDAGENS COMPUTACIONAIS PARA BIG DATA APLICADAS AO MERCADO DE AÇÕES: UM MAPEAMENTO SISTEMÁTICO

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Software da Universidade Tecnológica Federal do Paraná, como requisito parcial para a obtenção do título de Bacharel.

Orientador: Prof.ª Dra Érica Ferreira de Souza
Universidade Tecnológica Federal do Paraná

CORNÉLIO PROCÓPIO
2021

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Cornélio Procópio
Departamento Acadêmico de Computação
Engenharia de Software

TERMO DE APROVAÇÃO

Abordagens computacionais para Big Data aplicadas ao mercado de ações: um mapeamento sistemático

por

Nathiely Laiane Moraes Macedo

Este Trabalho de Conclusão de Curso de graduação foi julgado adequado para obtenção do Título de Bacharel em Engenharia de Software e aprovado em sua forma final pelo Programa de Graduação em Engenharia de Software da Universidade Tecnológica Federal do Paraná.

Cornélio Procópio, 27/08/2021

Professora Drª. Érica Ferreira de Souza

Prof. Titulação. Nome Professor Orientador

Professor Dr. Giovani Volnei Meinerz

Prof. Titulação. Nome Professor membro da banca

Professora Drª. Katia Romero Felizardo

Prof. Titulação. Nome professor membro da banca

"A Folha de Aprovação assinada encontra-se na Coordenação do Curso"

Aos meus pais pelo incentivo, perseverança e desafios enfrentados pelo meu sucesso.

AGRADECIMENTOS

Expresso minha gratidão as pessoas que estiveram presentes na minha trajetória até aqui.

Aos meus pais, pelo apoio, pelos sacrifícios, pelo amor e por acreditarem em mim.

A toda minha família, por acreditarem na minha capacidade e apoiarem minhas escolhas.

As minhas colegas de república, Loryane e Gabrielle, pelo companheirismo e amizade sincera.

Ao José Alexandre, pelo apoio e ferramentas fornecidas para a realização deste trabalho, os quais foram indispensáveis.

A minha orientadora, professora Érica Ferreira de Souza, pelo compromisso com este trabalho, pela paciência, pelas sugestões e disposição em contribuir com a pesquisa.

A função da educação é ensinar a pensar intensamente e pensar criticamente. Inteligência mais caráter: esse é o objetivo da verdadeira educação (Martin Luther King).

RESUMO

MACEDO, Nathiely Laiane Moraes. Abordagens computacionais para Big Data aplicadas ao mercado de ações: um mapeamento sistemático. 2021. 41 f. Trabalho de Conclusão de Curso – Curso de Engenharia de Software, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2021.

Contexto: Um dos maiores problemas no processo de tomada de decisão nas organizações é a baixa taxa de reutilização de conhecimento. Isso ocorre porque a maior parte do conhecimento nas organizações não é tratada, tornando difícil a sua articulação. É diante deste contexto que a Gestão do Conhecimento (GC) emerge como um importante meio para gerir o conhecimento útil. O conhecimento gerado a partir de discussões pode se tornar um item valioso. No setor financeiro, por exemplo, pode auxiliar investidores em sua tomada de decisão acerca das ações de determinadas empresas, negociadas em uma bolsa de valores, que em um determinado momento estejam mais atrativas. Neste setor, o volume de dados, a velocidade de geração e processamento dos dados de diferentes fontes criam desafios, tais como o armazenamento, processamento, visualização e, principalmente análise dos dados. O grande volume de dados produzidos por diferentes fontes, distribuídas e descentralizadas que geram rapidamente dados com relações complexas e em evolução é chamado *Big Data*. A análise desse grande volume de dados tem como objetivo extrair informações a respeito do domínio e o resultado dessa análise pode auxiliar as organizações na tomada de decisões. **Objetivo:** O objetivo deste trabalho é sumarizar as principais abordagens computacionais para *Big Data* aplicadas no mercado de ações a fim de identificar o estado da arte na área, bem como pesquisas futuras. **Método:** Um mapeamento sistemático foi conduzido. Um mapeamento sistemático fornece uma visão ampla de uma área de pesquisa, para determinar se há evidências de pesquisa sobre um determinado tópico. **Resultados:** Foram selecionados 115 estudos relevantes para responder as questões de pesquisa. Os resultados mostraram que o tópico de pesquisa é recente e possui uma grande variedade de abordagens para *Big Data* aplicadas no mercado de ações. A análise sentimental e a utilização de séries temporais, por exemplo, são comumente aplicadas entre os estudos. **Conclusão:** De acordo com os resultados obtidos é possível afirmar que os pesquisadores estão investindo na criação de novas ferramentas, técnicas, algoritmos e modelos para lidar com grande conjuntos de dados em tempo real e históricos do mercado de ações. Os modelos baseados em redes neurais artificiais e algoritmos de aprendizado de máquina são recorrentes nas pesquisas. Foi constatado que é possível criar abordagens para análises de previsão no mercado de ações utilizando fontes de dados de diferentes naturezas, como mídias sociais e relatórios de preços históricos. Espera-se que os resultados alcançados por esta pesquisa possam fornecer uma orientação para posicionar apropriadamente novas atividades de pesquisa no tópico investigado.

Palavras-chave: Gestão do Conhecimento. Big Data. Mapeamento Sistemático.

ABSTRACT

MACEDO, Nathiely Laiane Moraes. Big Data Computational Approaches Applied to the Stock Market: a systematic mapping. 2021. 41 f. Trabalho de Conclusão de Curso – Curso de Engenharia de Software, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2021.

Context: One of the biggest problems in the decision-making process in organizations is the low rate of knowledge reuse. It this occurs because most of the knowledge in organizations is not addressed, making articulation difficult. So, in this context the Knowledge Management (KM) emerges as an important means to manage useful knowledge. Knowledge generated from discussions can become a valuable item. In the financial sector, for example, it can help investors in their decision-making about the actions of certain companies, which are traded on a stock exchange, which at a given moment are more attractive. In this sector, the data volume, the speed of generation and data process of from different sources create challenges, such as storage, processing, visualization and, mainly, data analysis. The large data volume produced by different distributed and decentralized sources that quickly generate data with complex and evolving relationships is called *Big Data*. The analysis of this large data volume aims to extract information about the domain and the result of this analysis can help organizations in decision making. **Goal:** The objective of this work is to summarize the main approaches of *Big Data* applied in the financial market in order to identify the state of the art in the area, as well as future research. **Method:** An systematic mapping was conducted. A systematic mapping provides a broad view of a research area to determine if there is research evidence about a particular topic. **Results:** 115 relevant studies were selected to answer the research questions. The results showed that the research topic is recent and has a wide variety of approaches to Big Data applied to the stock market. Sentimental analysis and the use of time series, for example, are commonly applied across studies. **Conclusion**: According to the results obtained, it is possible to affirm that researchers are investing in the creation of new tools, techniques, algorithms and models to deal with large sets of real-time and historical stock market data. Models based on artificial neural networks and machine learning algorithms are recurrent in research. It was found that it is possible to create approaches for forecast analysis in the stock market using data sources of different natures, such as social media and historical price reports. It is hoped that the results achieved by this research can provide guidance to appropriately position new research activities on the investigated topic.

Keywords: Knowledge Management. Big Data. Systematic Mapping.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BCC	*British Broadcasting Corporation*
BTC	Bitcoin
CDB	Certificado de Depósito Bancário
CE	Critério de Exclusão
CI	Critério de Inclusão
CSI 300	*China Securities Index*
CNN	*Cable News Network*
DJIA	*Dow Jones Industrial Average*
ETF	*Exchange-traded Fund*
EUA	Estados Unidos da América
FTSE	*Financial Times Stock Exchange*
GARCH	*Generalized Autoregressive Conditional Heteroskedasticity*
GC	Gestão do Conhecimento
IEEE	*Institute of Electrical and Electronics Engineers*
IPO	Oferta Pública Inicial de Ações
JS	*JavaScript*
JSON	*Javascript Object Notation*
KM	*Knowledge Management*
LSTM	*Long Short-Term Memory*
MS	Mapeamento Sistemátic]
MLP	*Multilayer Perceptron*
NASDAQ	*National Association of Securities Dealers Automated Quotations*
NoSQL	*Not Only SQL*
NSE	*NationalStock Exchange of India*

NYSE	*New York Stock Exchange*
PCA	*Principal Component Analysis*
PNL	Processamento de Linguagem Natural
QP	Questão de Pesquisa
RBF	*Radial Basis Function*
RDB	Recibo de Depósito Bancário]
RF	*Random Florest*
RNN	*Recurrent Neural Network*
RPN	*Rough Petri Nets*
RS	Revisão Sistemática
S&P500	*Standard & Poor's 500*
SGBD	Sistemas de Gerenciamento de Banco de Dados
SSE	Shanghai Stock Exchange Composite Index
SVM	*Support Vector Machine*
USO	*United States Oil Fund*

SUMÁRIO

1 INTRODUÇÃO

Um dos maiores problemas no processo de tomada de decisão nas organizações é a baixa taxa de reutilização de conhecimento. Isso ocorre porque a maior parte do conhecimento nas organizações não é tratada, tornando difícil sua articulação. Diante deste contexto, a Gestão do Conhecimento (GC) emerge como um importante meio para gerir o conhecimento útil (DAVENPORT; PRUSAK, 2000). A GC é uma estratégia deliberada e sistemática de otimização de negócios que seleciona, destila, armazena, organiza, empacota e comunica informações essenciais aos negócios de uma empresa, de maneira a melhorar o desempenho do funcionário e a competitividade corporativa (NONAKA; TAKEUCHI, 1995).

O principal objetivo da GC é promover o armazenamento e o compartilhamento de conhecimento, bem como o surgimento de novos conhecimentos (O'LEARY; STUDER, 2001). O conhecimento pode ser classificado em dois tipos: tácito e explícito (NONAKA; KROGH, 2009). O conhecimento tácito surge de experiências individuais e envolve fatores como habilidade, crença pessoal, perspectivas e valores. O conhecimento explícito, por sua vez, pode ser expresso de uma forma estruturada, tais como tabelas, figuras, representações, esquemas e diagramas.

A transformação do conhecimento tácito em explícito para que esse possa ser compartilhado como item de conhecimento e difundido, tem sido objeto de muito investimento por diferentes segmentos de organizações, resultando em tentativas práticas que focalizam na passagem do individual e pessoal para o coletivo/grupal (NONAKA; TAKEUCHI, 1995). No setor financeiro, por exemplo, um item de conhecimento pode auxiliar investidores em sua tomada de decisão acerca das ações de determinadas empresas, negociadas em uma bolsa de valores, que em um determinado momento estejam mais atrativas. Neste setor, o volume de dados, a velocidade de geração e processamento dos dados de diferentes fontes criam desafios, tais como o armazenamento, processamento, visualização e, principalmente análise dos dados.

O grande volume de dados produzidos por diferentes fontes autônomas, distribuídas e descentralizadas que geram rapidamente dados com relações complexas e em evolução é chamado *Big Data* (MARQUESONE, 2016). *Big Data* refere-se a grandes conjuntos de dados impossíveis de serem gerenciados e processados usando apenas ferramentas tradicionais de gerenciamento de dados, exigindo plataformas computacionais mais complexas para serem analisados (AKOKA; COMYN-WATTIAU; LAOUFI, 2017).

Na literatura, tanto pesquisas sobre *Big Data*, como mercado financeiro tem aumentado consideravelmente. No mercado financeiro, *Big Data* passou a ser utilizado em várias empresas que buscam criar rotinas mais eficazes, avaliar com mais velocidade as tendências de mercado e rastrear fraudes com maior precisão. Dessa forma, a análise dos dados gerados passou a ser uma ferramenta estratégica, tornando a empresa mais preparada para lidar com mudanças no mercado e com os desafios do mundo digital.

1.1 OBJETIVOS

O objetivo deste trabalho é sumarizar as principais abordagens de *Big Data* aplicadas no mercado financeiro a fim de identificar o estado da arte na área. Para alcançar o objetivo proposto, um mapeamento sistemático da literatura é conduzido. Um mapeamento sistemático fornece uma visão ampla de uma área de pesquisa para determinar se há evidências de pesquisa sobre um determinado tópico (KITCHENHAM; CHARTERS, 2007).

1.2 ORGANIZAÇÃO DO TRABALHO

O presente trabalho está organizado da seguinte forma: no Capítulo 2 são apresentados os principais conceitos os trabalhos relacionados com o objetivo do projeto. No Capítulo 3 é apresentada a metodologia utilizada. No Capítulo 4 são apresentados a extração dos dados e os resultados. No Capítulo 5, a conclusão deste trabalho é apresentada.

2 REVISÃO DE LITERATURA

Neste capítulo são apresentados brevemente os principais conceitos relacionados ao estudo. Na Seção 2.1 são apresentados conceitos relacionados à Gestão do Conhecimento (GC). Na Seção 2.2 são apresentados os conceitos sobre *Big Data*. Na Seção 2.3 são apresentados os conceitos sobre Aprendizado de Máquina. A Seção 2.4 descreve os principais conceitos sobre Mercado Financeiro. Na Seção 2.5 é descrito o processo para condução de um mapeamento sistemático. Por fim, na Seção 2.6 são apresentados alguns trabalhos relacionados.

2.1 GESTÃO DO CONHECIMENTO

A popularização da GC nas organizações se tornou mais evidente a partir de 1980, quando emerge como uma solução para se obter conhecimento útil baseado em um grande volume de informações. O gerenciamento dessas informações, que podem ser vantajosas diante da competitividade do mercado, exige novos mecanismos, produtos e processos organizacionais que comportem flexibilidade durante o tratamento (SPENDER, 1996).

O conhecimento é visto como um item valioso quando combinado com experiências dos funcionários da organização e agregado à tecnologias. Várias organizações em diferentes segmentos estão prestando atenção ao conhecimento gerado e gerenciando ativamente seu capital intelectual (DAVENPORT; PRUSAK, 2000). A GC proporciona a organização do conhecimento, tornando-o acessível para que sejam armazenados e compartilhados.

Para Probst, Raub e Romhardt (2002), a GC pode ser definida como um processo que quando utilizado por organizações obtêm valor a seus ativos intelectuais. Os valores intelectuais são obtidos a partir de relações organizacionais e para gerá-los é necessário que os conhecimentos sejam disseminados entre setores, funcionários e até concorrentes. De acordo com Davenport e Prusak (2000), a GC trata de um processo de identificação, captura, compreensão, manutenção, criação, compartilhamento e utilização do conhecimento em uma organização.

A GC pode ser fundamentada em três conceitos básicos: dado, informação e conhecimento (DAVENPORT; PRUSAK, 2000). Os dados são fatos objetivos que sozinhos não possuem significado, são eventos distintos. Informação é o que sucede os dados, trata-se da interpretação e contextualização dos dados, atribuindo significado. Por fim, o conhecimento é o conjunto de informações que implicam na sabedoria, com valores atribuídos. No Quadro 1 é apresentada a diferenciação de dado, informação e conhecimento definida por Davenport e Prusak (2000).

O conhecimento nas organizações também pode ser classificado em dois tipos: tácitos e explícitos (NONAKA; TAKEUCHI, 1995). O conhecimento tácito trata de um conhecimento subjetivo, que pode estar na mente das pessoas ou em comportamentos, tornando difícil conceituar, o que o torna mais valioso, já que demanda esforço para ser capturado e pode

Quadro 1 – Diferenciação Entre Dado, Informação e Conhecimento

DADO	**INFORMAÇÃO**	**CONHECIMENTO**
Simples observação sobre o estado do mundo	Dados dotados de relevância e propósito	Informação valiosa da mente humana. Inclui reflexão, síntese e contexto.
Facilmente estruturado	Requer unidade de análise	Difícil estruturação.
Facilmente obtido por máquinas	Exige consenso em relação ao significado	Difícil captura em máquinas
Frequentemente quantificado	Exige necessariamente a medição humana	Frequentemente tácito
Facilmente transferível	-	Difícil transferência

Fonte: Davenport e Prusak (2000)

conter muita informação útil. Já o conhecimento explícito é naturalmente tangível, se trata de um conhecimento acessível, disponibilizado em alguma forma de mídia ou texto (DALKIR, 2005).

A extração de diferentes tipos de conhecimento em uma organização requer a exteriorização do conhecimento, definida como a conversão de conhecimentos tácitos em explícitos com o objetivo de compartilhá-los como itens de conhecimento (NONAKA; TAKEUCHI, 1995). Essa transformação requer metodologias que adaptem as informações individuais para um contexto coletivo (DAVENPORT; PRUSAK, 2000; NONAKA; KROGH, 2009). É possível encontrar na literatura diferentes ciclos de GC, também conhecidos como modelos, com esse propósito. Um ciclo de GC possui atividades que permitem a manipulação do conhecimento, por exemplo, criação, retenção, reutilização, compartilhamento e aplicação do conhecimento, seja ele organizacional, inter-organizações, grupal ou individual. Além dos ciclos, também existem diversas práticas que podem ser utilizadas para propagar o conhecimento gerado pelas organizações, tais como lições aprendidas, páginas amarelas, memória organizacional e base de conhecimento (DAVENPORT; PRUSAK, 2000; PROBST; RAUB; ROMHARDT, 2002; BATISTA, 2006).

Aurum, Daneshgar e Ward (2008) defendem a GC como uma ferramenta que auxilia a capacidade de uma organização de se estruturar diante do seu próprio ambiente, como uma estratégia de negócio que incorpora conhecimento em seus processos. O grande desafio dessa estratégia é filtrar os conhecimentos relevantes para a empresa, evitando gastos e investimentos desnecessários com conhecimentos que não são do interesse da organização (BUKOWITZ; WILLIAMS, 2000; RODRIGUEZ-ELIAS et al., 2008; ESTEVES, 2017). Além disso, em muitas organizações, o grande volume de dados gerados, a velocidade de geração e o processamento dos dados de diferentes fontes, criam desafios, tais como o armazenamento, processamento, visualização e, principalmente análise dos dados. É nesse contexto que o investimento em *Big Data* torna-se realidade de muitas empresas.

2.2 BIG DATA

Os avanços tecnológicos e a ascensão da Internet influenciaram o crescimento do volume de dados gerados pela população e pelas máquinas, fazendo com que as pessoas adaptem seu cotidiano com a implantação de tecnologias digitais em suas tarefas habituais. Para Marquesone (2016), o uso de dispositivos móveis é um dos principais motivos para esse crescimento exorbitante de dados, pois além de auxiliar nas tarefas pessoais, permite o compartilhamento em tempo real de conhecimento, que rapidamente é espalhado sendo compartilhado muitas vezes. Outros dois motivos citados como principais causas do crescimento do volume dos dados são o aumento do poder de processamento dos computadores e a diminuição do custo de armazenamento.

Diante desse contexto, tem-se a necessidade de um modelo de banco de dados que suporte um volume muito grande de dados, capazes de serem sintetizados a fim de obter informações que possuem valor organizacional e econômico. Apesar da efetiva recuperação de dados apresentada pelos bancos de dados relacionais, o modelo relacional é ineficaz para lidar com a grande escala e diferentes configurações de dados (ZAFAR et al., 2016). O *Big Data* é uma tecnologia que pode solucionar esse problema e vem sendo utilizado em diversos setores.

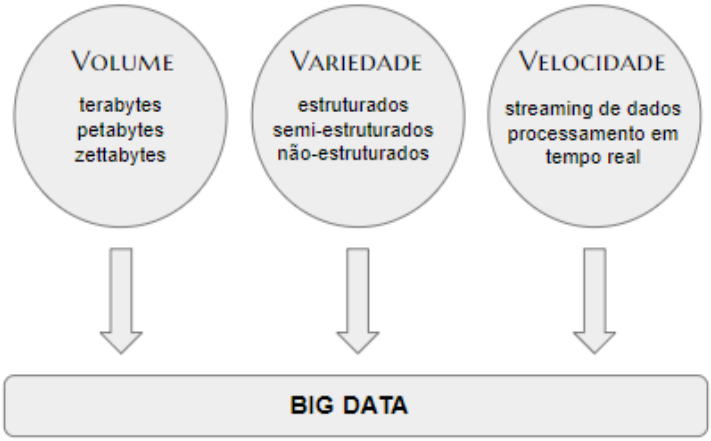
O *Big Data* é definido como um banco de dados de alto volume, alta velocidade e/ou alta variedade de dados que exigem novos paradigmas de processamento para permitir a descoberta de ideias (BEYER; LANEY, 2012). O tamanho típico dos dados é consideravelmente grande, isto é, na magnitude de Petabytes ou Exabytes, e está aumentando rapidamente a cada ano (volume). À medida que mais pessoas têm acesso à Internet, a taxa de dados disponíveis também aumenta (velocidade). Além disso, os dados podem se originar de diferentes fontes e conter vários elementos, como imagens, vídeos e mensagens de voz (variedade) (L'HEUREUX et al., 2017), caracterizando os 3 V's do *Big Data*. Os 3 V's do *Big Data* são explicados por Beyer e Laney (2012) da seguinte forma:

- Volume: está relacionado ao fluxo contínuo de dados gerados e o volume acumulado desses dados;
- Velocidade: diz respeito a velocidade em que os dados gerados precisam ser armazenados, processados e disponibilizados para outros objetivos;
- Variedade: está relacionada à grande variedade de formatos e estruturas de dados.

O projeto *Big Data* é composto por diferentes fases. A Figura 2 apresenta as fases de um *Big Data*. A seguir são apresentadas brevemente as características de cada uma das fases.

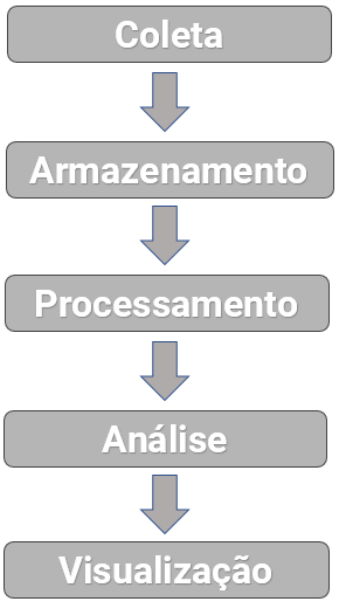
- Coleta: o objetivo da coleta é adquirir dados que possam ser transformados em informações relevantes para o projeto. Nesta fase, os dados são classificados segundo sua fonte (interna ou externa) e formato (estruturado, semi-estruturado e não-estruturado).
  As fontes internas compreendem todos os dados adquiridos dentro da organização, seja oriundo de suas relações interpessoais, de máquinas ou de documentos empresariais. As fontes internas são importantes para o conhecimento sobre a empresa e podem auxiliar na

Figura 1 – Os 3 V's de Big Data



Fonte: Marquesone (2016)

Figura 2 – Fases do Projeto Big Data



Fonte: Marquesone (2016)

tomada de decisões (MARQUESONE, 2016). Os dados externos são de domínio público mas estão na área de interesse das soluções.

Quanto ao formato, a *Education Services* EMC (2015) define os dados estruturados como aqueles que possuem estrutura, tipo e formato bem definido, como as planilhas, e os dados não-estruturados como os que não possuem nenhum tipo de estrutura definida, como arquivos de áudio. Marquesone (2016) caracteriza como dados semi-estruturados os que utilizam marcadores e hierarquias, sem possuir uma estrutura fixa, é o caso dos arquivos JSON (Javascript Object Notation), por exemplo.

- Armazenamento: Os modelos de banco de dados relacionais são insuficientes para as novas tecnologias, como as redes sociais e os aplicativos mobile (TANG; FAN, 2016). O grande volume de dados gerados em tempo real demanda maior capacidade de leitura e escrita e abordagens diferente das soluções estruturadas que possam armazenar diferentes formatos de dados (CHEN; ZHANG, 2014). O armazenamento de dados semi-estruturados ou não estruturados exigem uma estrutura flexível, diante do fato de que torna-se inviável o conhecimento antecipado do formato de todos os dados (MARQUESONE, 2016).
  O termo NoSQL(*Not Only SQL*) é a terminologia para os novos modelos de armazenamento de dados que surgiram devido a essa ineficiência (ZAFAR et al., 2016). Assim, os modelos NoSQL podem ser classificados de acordo com a estrutura de armazenamento dos dados. Atualmente, existem quatro modelos principais de SGBD (Sistemas de Gerenciamento de Banco de Dados): modelo orientado a chave-valor, orientado a documentos, orientado a colunas e orientado a grafos.
- Processamento: No processamento dos dados, é importante manter a escalabilidade da solução. Para Marquesone (2016), uma solução é escalável se for capaz de manter o desempenho desejável mesmo com a adição de nova carga. Com o ágil crescimento exponencial dos dados, alcançar essa escalabilidade é um grande desafio.
  O processamento de grande volume de dados pode ser realizado a partir do processamento em lote, que é indicado quando a solução deve processar um conjunto volumoso de dados já coletados, ou o processamento em tempo real, que processa os dados de maneira imediata, isto é, quando são gerados.
  Para aumentar a velocidade do processamento dos dados é utilizada a técnica de processamento distribuído. Existem ferramentas, como o *Hadoop*, que realiza o processamento distribuído combinando os processamentos em lote e tempo real.
- Análise: Antes de realizar a análise é necessário aplicar técnicas de limpeza dos dados, que consistem em remover conteúdos que não são relevantes para a análise (BERNARDO; HENRIQUES; LOBO, 2017). Após o tratamento, os dados devem ser analisados para extrair informações sobre o domínio da solução a partir da detecção de grupos, identificação de padrões e da classificação dos dados.
  Para criar um modelo de análises, deve-se obter conhecimento sobre o domínio, a fim de identificar os objetivos da análise e as questões a serem respondidas, além do entendimento das estruturas e relações dos dados. Os dados devem ser preparados com a técnica de limpeza para que sejam modelados, definindo um modelo de análise. A precisão do modelo é medida a partir dos resultados obtidos e, finalmente utilizado sob monitoração com o objetivo de manter a sua confiabilidade.
- Visualização: A visualização dos dados é realizada através de análises exploratórias e explanatórias, que acontecem antes, durante e depois da análise dos dados e compreende etapas que vão desde a captura dos dados até a representação gráfica. Para isto, é necessário desenvolver uma interface que atenda o propósito específico do projeto de *Big*

Data em questão, apresentando informações coerentes e compreensíveis (MARQUESONE, 2016). Existem várias ferramentas desenvolvidas para a visualização de grandes volumes de dados que constroem representações gráficas dos resultados com recursos interativos.

De acordo com Marquesone (2016), o *Big Data* já é aplicado nas áreas governamentais, no setor financeiro, na área de transporte e automação, no setor de varejo, em áreas de marketing e na área de seguros. No setor financeiro, em especial, a pesquisa tem aumentado consideravelmente. No mercado financeiro, *Big Data* passou a ser utilizado em várias empresas que buscam criar rotinas mais eficazes, avaliar com mais velocidade as tendências de mercado e rastrear fraudes com maior precisão. A análise dos dados relacionados com bolsa de valores e mercado de ações, por exemplo, passaram a ser uma ferramenta estratégica, tornando a empresa mais preparada para lidar com mudanças no mercado e com os desafios do mundo digital.

2.3 APRENDIZADO DE MÁQUINA

O aprendizado de máquina é uma coleção de métodos que permitem que os computadores automatizem a criação e a programação de modelos orientados a dados por meio de uma descoberta sistemática de padrões estatisticamente significativos nos dados disponíveis (BHAVSAR et al., 2017), para isto, aplica métodos estatísticos a fim de melhorar o desempenho com base em experiências anteriores ou na detecção de novos padrões presentes em grandes quantidades de dados (THEOBALD, 2017). A utilização do aprendizado de máquina permite lidar com tarefas complexas demais para serem resolvidas com algoritmos projetados por seres humanos (GOODFELLOW; BENGIO; COURVILLE, 2016).

O aprendizado de máquina é uma subárea da Inteligência Artificial que tem como objetivo a criação de máquinas inteligentes com a capacidade de aprender a resolver problemas tomando decisões sobre um conjunto de dados a fim de gerar resultados precisos. Pode ser classificado de acordo com o tipo de aprendizado, que pode ser supervisionado, não supervisionado ou semi-supervisionado;

- Supervisionado: Um conjunto de exemplos de treinamento com as respostas corretas (metas) é fornecido e, com base nesse conjunto de treinamento, o algoritmo generaliza para responder corretamente a todas as entradas possíveis (MARSLAND, 2014). O sistema é treinado para reconhecer padrões baseado em suas próprias experiências obtidas com o conjunto de dados rotulados, os algoritmos de aprendizado supervisionado são separados em duas categorias, classificação e regressão, que são utilizados de acordo com o tipo de dado.
- Não supervisionado: Depende apenas dos dados não rotulados subjacentes para identificar padrões ocultos de dados em vez de inferir modelos para pares de entradas e saídas conhecidas (BHAVSAR et al., 2017). Neste tipo de abordagem, o treinamento da máquina não é orientado, os padrões devem ser reconhecidos através de agrupamentos classificados de acordo com as semelhanças encontradas nos dados. Os algoritmos utilizados no

aprendizado não supervisionado se categorizam em *clusters* e associação.

- Semi-supervisionado: Os dados rotulados são custosos, pois precisam de interferência humana para gerar os conjuntos de treinamento, por outro lado, os dados não rotulados são fáceis de ser coletados mas exigem maior esforço para serem analisados, para resolver estes problemas, o aprendizado semi-supervisionado combina os dados rotulados e não rotulados a fim gerar classificadores mais eficientes. Dessa forma, os dados não rotulados são utilizados para modificar ou re-priorizar hipóteses obtidas apenas a partir de dados rotulados (ZHU, 2005).

2.4 MERCADO FINANCEIRO

O mercado financeiro é responsável por promover investimentos e o crescimento da economia através de interações entre tomadores e poupadores de recursos (NETO, 2012). O ambiente em que produtos financeiros são negociados é o mercado financeiro e as transações são realizadas a partir de investidores, intermediários e emissores;

- Investidores: Financiam projetos ou investimentos dos emissores emprestando dinheiro em trocas de juros futuros ou adquirindo participação societária. Compreendem pessoas físicas, fundos de investimentos ou empresas.
- Intermediários: Responsáveis por regulamentar e fiscalizar o mercado financeiro, supervisionando a relação entre investidores e emissores. São exemplos de intermediários os bancos de investimentos e as corretoras.
- Emissores: Empresas ou instituições que emitem títulos de valores mobiliários ou créditos, como ações de empresas de capital aberto ou títulos públicos.

Os produtos financeiros são os ativos que os envolvidos compram e vendem no mercado, podendo ser ações, títulos públicos, *Exchange-traded Fund* (ETF), debêntures, Certificado de Depósito Bancário (CDB), Recibo de Depósito Bancário (RDB), fundos imobiliários, etc. O mercado financeiro é divido de acordo com suas mercadorias, ramificando em mercado de créditos, mercado monetário, mercado de capitais e mercado de câmbios. O Quadro 2 exemplifica suas características.

Quadro 2 – Estruturas dos Mercados Financeiros

MERCADOS	**ATUAÇÃO**	**MATURIDADE**
Monetário	Controle dos meios de pagamentos da economia (liquidez)	Curtíssimo e curto prazos
Crédito	Créditos para consumo e capital de giro das empresas	Curto e médios prazos
Capitais	Investimentos, financiamentos e outras operações	Médio e longo prazos
Cambial	Conversão de Moedas	A vista e curto prazo

Fonte: Neto (2012)

Os mercados podem ser divididos ainda em primário e secundário (FORTUNA, 1998). No mercado primário ocorrem as primeiras negociações dos valores das ações emitidas pelas companhias, a fim de utilizar seus recursos para financiar seus investimentos. Já o mercado secundário é gerenciado pelas bolsas de valores, registrando somente a transação de propriedade dos títulos e valores mobiliários, sem determinar variações diretas no fluxo de recursos das instituições. O investidor tem a possibilidade de vender as ações adquiridas a fim de reaver o capital aplicado (NETO, 2012).

2.4.1 MERCADO DE AÇÕES

O mercado de ações é responsável por manipular as variações de ações dispostas para negociação. As ações são títulos que representam a menor fração do capital social de uma empresa que permitem aos portadores destes títulos participarem dos resultados da empresa (NETO, 2012; FORTUNA, 1998).

As ações podem ser classificadas em dois tipos, ordinárias e preferenciais (FORTUNA, 1998). As ações ordinárias qualificam o seu portador a participar de decisões na organização, garantindo direito a voto em discussões importantes da companhia, como a escolha de diretores e a destinação de lucros. As ações preferenciais garantem aos investidores prioridades no recebimento de dividendos e reembolso de capital caso a empresa sofra uma dissolução (NETO, 2012).

Os objetivos do mercado de ações variam de acordo com a perspectiva em que está inserido. Para as empresas, funciona como um meio de capitalizar seu lucro, extraindo recursos para o seu crescimento a partir do investimento de terceiros. Do ponto de vista da economia, as quedas ou elevações dos preços das ações são influenciados diretamente pelo cenário político e econômico dos países, ou seja, se a economia do país vai bem, as ações de suas empresas tendem a se valorizar (FORTUNA, 1998). Para os investidores, o mercado de ações é uma oportunidade de obter lucro e capital sem a necessidade de se envolver com a administração e o funcionamento das instituições acionárias.

As ações são disponibilizadas pelas empresas através de uma abertura de capital, a Oferta Pública Inicial de Ações (IPO), que consiste na primeira venda de ações no mercado por uma companhia (NETO, 2012). Quando uma organização abre uma IPO, passa a ser considerada uma sociedade anônima e suas ações se tornam visíveis para os investidores na bolsa de valores.

Com a aquisição das ações, os investidores assumem um risco decorrente de suas oscilações e a possibilidade de bonificação pelos seus títulos. Para obter lucro, os acionistas devem estar atentos e prever as mudanças na economia que podem influenciar a valorização de determinadas empresas. De acordo com Neto (2012), um bom investidor é aquele que sabe identificar o momento oportuno de adquirir uma ação no mercado, aproveitando para investir em ações com tendência a valorização antes dos outros investidores, a fim de vendê-las antes

da desvalorização.

2.4.2 BOLSA DE VALORES

A bolsa de valores gere o mercado secundário da economia, funcionando como uma organização de produtos financeiros. Para Neto (2012), as bolsas de valores são entidades com o objetivo de manter um local em condições adequadas para a realização, entre seus membros, de operações de compra e venda de títulos e valores mobiliários.

As transações da bolsa são realizadas no pregão, as informações sobre negócios que podem influenciar nos valores das ações são disponibilizadas para os membros do pregão. Tais membros são credenciados pelas corretoras, que negociam as ordens de compra e venda dos investidores (NETO, 2012). Fortuna (1998) define o pregão como o ambiente onde os operadores da bolsa de valores executam as ordens de compra e venda dadas pelos investidores às suas corretoras.

No Brasil, existe uma única bolsa de valores atuante, a "B3", que se consolidou em 2017 com a união da Cetip, Bovespa e BM&F com o objetivo de gerir todos os valores mobiliários de renda fixa e variável. Cada país possui sua(s) própria(s) bolsa(s) de valores e as oscilações de uma bolsa externa podem influenciar nos preços das ações de um determinado país, girando a economia mundial.

Para acompanhar o desempenho das ações, existem índices de bolsa de valores. No Brasil, o principal é o Índice da Bolsa de Valores de São Paulo, o Bovespa[1]. O Bovespa detém uma carteira das ações mais negociadas no país, desta forma, os investidores podem comparar seus papéis com os do índice para obter uma noção do quão rentável estão seus investimentos no mercado. O objetivo deste índice é refletir o desempenho e o perfil dos negócios que transitam os pregões, sendo apurado em tempo real de acordo com as ações realizados no mercado.

2.5 MAPEAMENTO SISTEMÁTICO

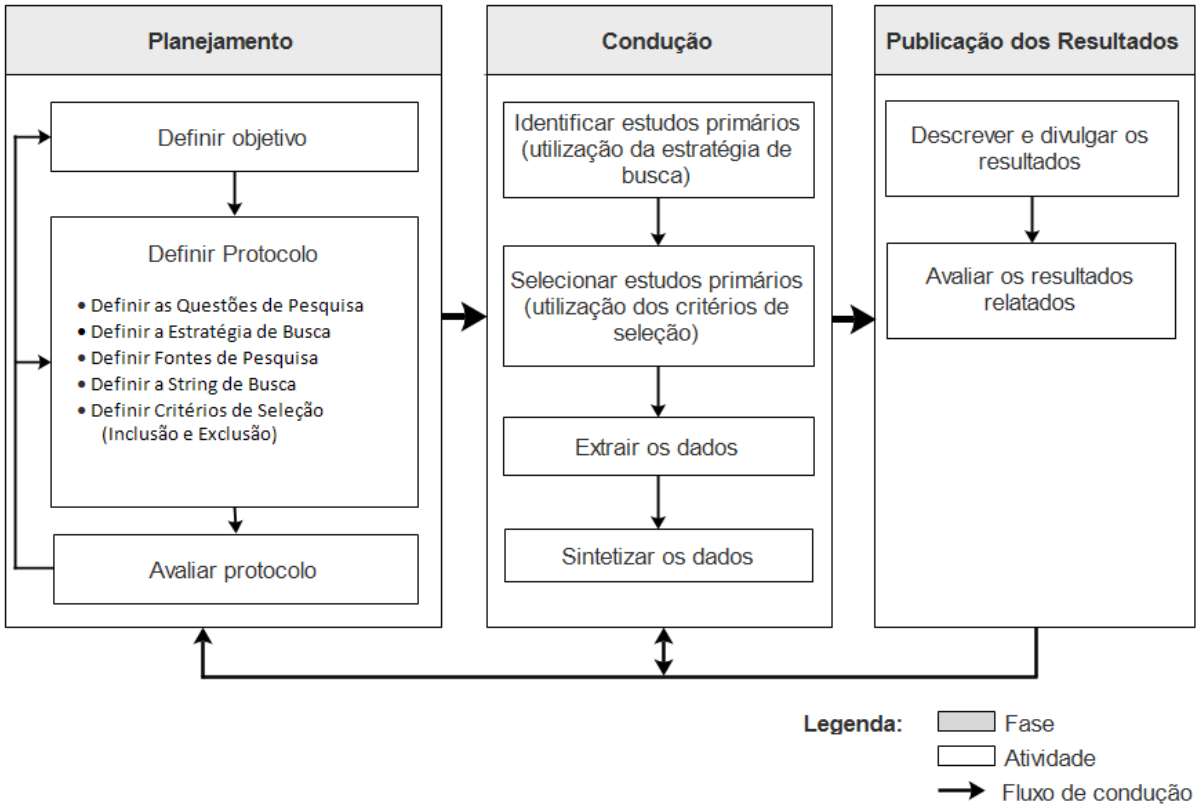
O objetivo da condução de estudos secundários é integrar e sintetizar evidências relacionadas a determinadas questões de pesquisa a partir de um conjunto de estudos chamados estudos primários (NAKAGAWA et al., 2017). As Revisões Sistemáticas (RS) e Mapeamentos Sistemáticos (MS) são tipos de estudos secundários que seguem um processo de pesquisa bem definido para identificar, analisar e interpretar tais evidências (KITCHENHAM; CHARTERS, 2007).

A diferença entre RSs e MSs é a amplitude da revisão. Ao contrário da RS, um MS é indicado quando não se têm uma visão muito definida sobre a área de pesquisa e se deseja apenas quantificar as evidências existentes. Na RS o pesquisador já conhece a área de pesquisa e o objetivo é aprofundar o estudo em um tópico específico. O MS possui questões de

[1] *www.bmfbovespa.com.br*

pesquisas mais genéricas que buscam explorar a área e compreender o domínio, enquanto uma RS possui questões de pesquisa mais específicas que buscam esclarecer um ponto específico dentro do domínio (PETERSEN; VAKKALANKA; KUZNIARZ, 2015). No entanto, por ambos serem estudos secundários, o processo para conduzir uma RS (Revisão Sistemática) ou um MS (Mapeamento Sistemático), envolve as mesmas atividades. O processo de uma RS e um MS envolve atividades de planejamento, condução e divulgação dos resultados. Esse é um processo fortemente iterativo e que pode requerer muitas revisões. A Figura 3 apresenta quais são as principais atividades que compõem cada uma dessas etapas.

Figura 3 – Processo de Condução de um Estudo Secundário



Fonte: Nakagawa et al. (2017)

A seguir cada etapa é descrita brevemente:

- **Planejamento:** No planejamento do processo de revisão deve ser identificada a motivação para a condução de um estudo secundário. Uma vez definidas as necessidades para conduzir o estudo, é definido um protocolo. Todos procedimentos para condução do estudo são definidos no protocolo. São definidos no protocolo a descrição dos objetivos, as questões de pesquisa a serem respondidas, a estratégia de busca, a *string* de busca, as fontes de pesquisa, os critérios de inclusão e exclusão, dentre outros.
  Após a definição do protocolo, é necessário que se avalie sua qualidade. Nesta atividade Kitchenham (2004) propõe que sejam testadas as *strings* de busca, validando se os

estudos relevantes para a pesquisa estão sendo encontrados. Pode-se utilizar um grupo de controle que contém estudos relevantes para a pesquisa indicado por um especialista da área. Além disso, um teste de viabilidade de execução da revisão deve ser conduzido, chamado teste piloto. Esse teste permite identificar modificações que sejam necessárias no protocolo.

- **Condução:** Na execução da revisão, o objetivo é encontrar estudos primários capazes de responder as questões de pesquisa definidas no protocolo. Esta busca pode acontecer de forma manual ou automática. A busca manual ocorre em anais de eventos ou revistas científicas sobre estudos que tratam um determinado assunto. A busca automática acontece na consulta de bases de dados eletrônicas (*IEEE Xplore, ACM, Springer, Science Direct*) por meio da *string* de busca definida.

  A partir dos estudos coletados nas bases de dados, o pesquisador realiza a seleção dos estudos a partir da aplicação dos critérios de inclusão e exclusão. Primeiramente é realizada uma seleção inicial em que se aplica os critérios de inclusão e exclusão apenas no título, resumo e palavras-chave. Caso o pesquisador identifique que o estudo não se enquadra nos critérios de busca, ele é excluído. Em seguida é conduzida uma seleção com o pesquisador lendo na íntegra os estudos selecionados anteriormente e novamente os critérios de inclusão e exclusão são aplicados. Ao final do processo de seleção uma amostra dos estudos é revisada para garantir que os critérios de inclusão e exclusão foram aplicados de maneira correta.

  Após a atividade de seleção, os dados contidos nos estudos primários devem ser extraídos e sumarizados a fim de responder as questões de pesquisa. Formulários de extração de dados podem ser utilizados para coletar os dados que sejam necessários e facilitar as análises e síntese dos resultados.
- **Publicação dos Resultados:** A última etapa está relacionada à escrita e divulgação dos resultados para os potenciais interessados. Isso pode ser realizado através da publicação por meio de relatórios técnicos, artigos de revistas ou conferências, capítulos de livros e trabalhos de conclusão de curso.

2.6 TRABALHOS RELACIONADOS

Com o interesse pelas empresas em se trabalhar com *Big Data*, houve um aumento considerável em pesquisas nesta área. Diversos estudos secundários tem sido publicados nesse sentido em diversas áreas, bem como no setor financeiro, escopo deste projeto. A seguir alguns desses estudos são apresentados brevemente.

Em Akoka, Comyn-Wattiau e Laoufi (2017), foi conduzido um mapeamento sistemático a fim de analisar como os pesquisadores tem compreendido o conceito de *Big Data*. Para isso algumas perguntas são investigadas nos estudos selecionados, tais como: Qual é a tendência anual das publicações? Quais são os tópicos mais investidos em *Big Data*? Por que a pesquisa é realizada? O que a pesquisa em *Big Data* produz?, dentre outras. Os resultados do trabalho

conduzido por Akoka, Comyn-Wattiau e Laoufi (2017) mostram que as principais contribuições em *Big Data* foram feitas pela comunidade de científica, atestadas por um aumento considerável de publicações científicas que abordam o tema. Os pesquisadores estão cada vez mais envolvidos em pesquisas que combinam *Big Data* e *Analytics*, *Cloud*, Internet das coisas, mobilidade ou mídias sociais. Aspectos de segurança e qualidade de *Big Data* também tem sido considerados importantes nas pesquisas.

Bhardwaj e Singh (2017) conduziram um mapeamento sistemático com o objetivo de obter uma visão geral do estado da análise de *Big Data* nos processos de governança. Para alcançar o resultado, as principais questões levantadas para os estudos selecionados foram: Quais são as áreas de aplicação na governança? Quais são as ferramentas usadas na análise de *Big Data* na governança? Qual é a diferença entre governança com e sem análises de *Big Data*? Quantos artigos abrangem as diferentes categorias de área de aplicação e ferramentas usadas na análise de *Big Data* em governança? Como resultado, constata-se que a análise de *Big Data* é ativa no campo da governança, sendo a análise de redes inteligentes a principal aplicação. O estudo comprova que a governança com *Big Data* permite melhorar o gerenciamento de recursos através de predições, se destacando em relação a governança tradicional.

Em outro importante estudo, realizado por Laigner et al. (2018) foi conduzida uma Revisão Sistemática da Literatura com o propósito de verificar as pesquisas existentes em Engenharia de Software relacionadas com *Big Data* identificando suas contribuições e métodos de desenvolvimento. A revisão foi baseada em uma questão primária: quais tipos de abordagens de engenharia de software têm sido propostas para suportar o desenvolvimento de sistemas de *Big Data*?. Outras questões foram levantadas a fim de caracterizar métodos, objetivos, domínios de aplicação, tipos de contribuição, e outras características das pesquisas. Após a condução da revisão, observa-se que os estudos sobre abordagens de Engenharia de Software se concentram principalmente nas particularidades dos sistemas que envolvem *Big Data*, indicando um consenso entre indústria e academia de que as abordagens tradicionais não suportam as necessidades de desenvolvimento dos projetos de *Big Data* devido a sua complexidade e o volume dos dados. Dos 52 estudos selecionados, entre 2011 e 2016, a maioria se refere a identificação de lacunas em metodologias de desenvolvimento existentes no contexto de sistemas de *Big Data*. Há também uma alta incidência de estudos que visam encontrar soluções de arquitetura de software não convencionais que suportem aplicações que utilizam *Big Data*.

Por fim, em Bach et al. (2019) foi realizada uma Revisão Sistemática da Literatura sobre a mineração de texto para análises de *Big Data* no setor financeiro, a fim de responder três questões de pesquisa: Qual é o núcleo intelectual do campo? Quais técnicas são usadas no setor financeiro para mineração textual, especialmente na era da Internet, *Big Data* e mídias sociais?, e quais fontes de dados são as mais usadas para mineração de texto no setor financeiro e para quais fins?. As questões foram aplicadas em 123 artigos selecionados da base de dados *Web of Science*[2]. Os estudos mais relevantes estão focados na previsão de preços de

[2] *https://www.webofknowledge.com/*

ações, previsões de mercado e detecção de fraudes financeiras utilizando mineração de textos online. Ainda, foram identificadas várias técnicas de mineração de dados que auxiliam empresas e equipes de pesquisa a transformar informações em conhecimento, além de detectar que as fontes de dados internas são pouco utilizadas. A maioria das pesquisas utilizam fontes externas como notícias e posts de mídias *online*.

3 METODOLOGIA

Neste capítulo são apresentados os processos da metodologia utilizados.

3.1 DELINEAMENTO DA PESQUISA

O objetivo deste trabalho é sumarizar as principais abordagens de *Big Data* aplicadas no mercado de ações. Para alcançar o objetivo proposto, um mapeamento sistemático da literatura foi conduzido. Conforme apresentado na Figura 3, na etapa de planejamento é definido um protocolo a ser seguido, o qual define todos os procedimentos a serem realizados para condução do mapeamento.

A seguir é apresentado o protocolo utilizado para guiar o mapeamento sistemático conduzido neste trabalho:

A. Questões de Pesquisa (QP)

Esse mapeamento sistemático teve como objetivo responder as seguintes questões de pesquisa:

(QP1) Quando e onde os estudos foram publicados?

(QP2) Qual o propósito do estudo no mercado de ações?

(QP3) Quais são as fontes de dados utilizadas das quais os dados são coletados? (*Brazilian Depositary Receipts*- BDR), Fóruns, Wiki, tweets)?

(QP4) Quais técnicas, algoritmos, modelos e ferramentas foram empregados?

(QP5) Quais das etapas do *Big Data* são contempladas no artigo?

B. String de Busca

Na definição da *String* de busca são consideradas duas áreas - "*Big Data*" e "Mercado Financeiro" (Quadro 3).

Quadro 3 – String de busca - Aplicação de Big Data no Mercado Financeiro

Área	**Palavra-chave**
Big Data	*"Big data", "Data Science"*
Mercado Financeiro	*"Financial Market", "Stock Market", "Stock Price", "Stock data analysis", "stock movement", "financial big data"*
String de busca:	*("big data" OR "Data Science") AND ("Financial Market" OR "Stock Market" OR "Stock Price" OR "Stock data analysis" OR "stock movement" OR "financial big data")*

C. Base de Busca

A base de busca utilizada para identificar os estudos primários a partir da *string* de busca é a *Scopus*[1]. A base da *Scopus* é considerada a maior base de dados de resumos e citações de literatura revisada por pares, com mais de 60 milhões de registros. Além disso, a *Scopus* atribui artigos de outros editores internacionais, incluindo a *Cambridge University Press*, *Institute of Electrical and Electronics Engineers* (IEEE), *Nature Publishing Group*, *Springer*, *Wiley-Blackwell* e a *Elsevier*.

Os estudos retornados da base da *Scopus* foram catalogados e armazenados adequadamente em uma planilha. Esse catálogo apoia nos procedimentos de seleção e extração dos dados de cada estudo.

D. Critérios de Seleção

Os critérios de seleção são organizados em um Critério de Inclusão (CI) e sete Critérios de Exclusão (CE).

(CI1) O estudo deve apresentar abordagens de *Big Data* aplicadas no mercado financeiro.

(CE1) O estudo não possui um resumo;
(CE2) O estudo foi publicado somente como resumo;
(CE3) O estudo não está escrito em Inglês;
(CE4) O estudo é uma versão mais antiga de um estudo já considerado;
(CE5) O estudo não é um estudo primário; e
(CE6) O estudo não apresenta abordagens de *Big Data* aplicadas ao mercado financeiro.

E. Avaliação

Antes de conduzir o mapeamento, o protocolo foi testado. O teste foi realizado com o objetivo de verificar a viabilidade e adequação do protocolo, com base em um conjunto de estudos pré-selecionados e considerados relevantes para a investigação. Primeiramente, o processo de seleção foi conduzido pela aluna deste trabalho de conclusão de curso e, posteriormente, o orientador e mais um especialista da área realizam uma validação da seleção. Além disso, para definir e calibrar a *string* de busca foi considerado um grupo de estudos chamado grupo de controle.

No próximo capítulo são apresentados os resultados da execução das etapas de seleção considerando o protocolo definido, bem como os resultados da extração e síntese dos dados.

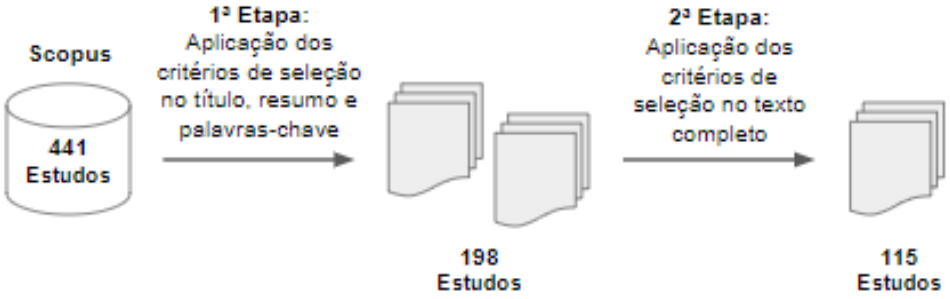
[1] *https://www.scopus.com/*

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Neste capítulo são apresentados a obtenção dos dados e os principais resultados da pesquisa. Na Seção 4.1 são apresentados as etapas de seleção dos estudos. As respostas das questões de pesquisa são discutidos na Seção 4.2.

4.1 SELEÇÃO DOS ESTUDOS

Figura 4 – Processo de Seleção



A string de busca (Quadro 3) foi executada na base da Scopus considerando o título, palavras-chave e resumo. Foram considerados estudos publicados de 2013 até setembro de 2019. Um total de 441 estudos foram retornados. O processo de seleção foi conduzido em duas etapas, como apresentado na Figura 4.

Na primeira etapa, o critério de inclusão e os critérios de exclusão CE1, CE3 e CE7 foram aplicados considerando apenas o título, o resumo e as palavras-chave. Foram eliminadas 244 publicações (aproximadamente 55%). O critério que excluiu o maior número de publicações foi o CE7, pois muitos dos estudos retornados possuem foco no mercado financeiro, mas não tratavam do mercado de ações especificamente.

Na segunda etapa, todos os critérios de inclusão e exclusão foram aplicados considerando o texto completo. Foram eliminados 83 estudos. Nesta etapa, o critério de exclusão que eliminou o maior número de artigos também foi o critério CE7, seguido pelo critério CE6, pois não foi possível obter acesso ao texto completo de 25 estudos. Como resultado, 115 estudos foram selecionados (aproximadamente 26% do total de artigos). A Tabela 1 apresenta a relação dos critérios de exclusão e o número de artigos excluídos. . A lista dos 115 artigos selecionados bem como a versão completa da extração de dados em cada etapa de seleção pode ser consultada na planilha de controle de estudo secundário[1].

Os 115 estudos selecionados foram analisados com o objetivo de responder as questões de pesquisa. As classificações dos artigos foram identificadas a partir das respostas das questões

[1] *https://bityli.com/Ch2PG*

Tabela 1 – Relação de Artigos Excluídos por Critério de Exclusão

Critério de Exclusão	**Etapa**	**Total de Excluídos**
CE1	1	1
	2	1
CE2	2	1
CE3	2	3
CE4	2	2
CE5	2	1
CE6	1	54
	2	25
CE7	1	177
	2	50

de pesquisa. Os estudos foram agrupados de acordo com os que possuem as mesmas classificações, definindo as categorias. O mesmo artigo pode ser mencionado em várias classificações diferentes, portanto a soma de ocorrências e porcentagens por classificação pode ultrapassar os 115 artigos, já que as quantidades são em relação ao número de artigos que compõem as categorias e representação em porcentagem da porção desses artigos em relação ao total. Além disso, em algumas QPs, classificações que tiveram poucas ocorrências são agrupados na categoria chamada "Outros".

4.2 RESULTADOS DO MAPEAMENTO

A seguir, são apresentados os principais resultados para cada QP definida na Seção 3.1.

QP1. Quando e onde os estudos foram publicados?

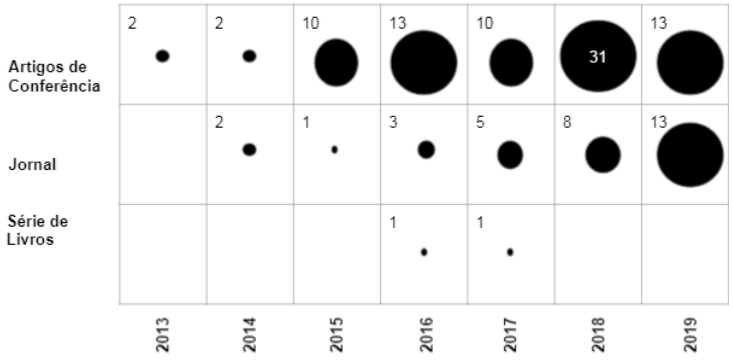
Três tipos de veículo de publicação foram encontrados para os estudos selecionados: artigos de conferências, revistas e séries de livros. A distribuição dos 115 estudos selecionados publicados até 2019 é mostrada na Figura 5.

O meio de publicação mais utilizado foram os artigos de conferência, totalizando 70% (81 estudos) das publicações. O segundo meio mais utilizado foram as revistas, com 27% das publicações (32 estudos). Já o menos utilizado foram as séries de livros, com apenas 1% das publicações, alcançando somente dois artigos durante os 7 anos contemplados por este estudo.

Entres os 81 estudos publicados em conferências, as conferências mais apresentadas foram a *IEEE International Conference on Big Data* e a *Advances in Intelligent Systems and Computing*. Já as revistas com o maior número de publicações entre os artigos selecionados foram a *Expert Systems with Applications*, a *Plos One* e a *Information Systems*.

Percebe-se que o campo de pesquisa de abordagens computacionais utilizando *Big Data* no mercado de ações é recente. O número de pesquisas tem um aumento considerável em 2018. Nos últimos dois anos contemplados por este trabalho, 2018 e 2019, o número de

Figura 5 – Distribuição dos Estudos Selecionados por Ano

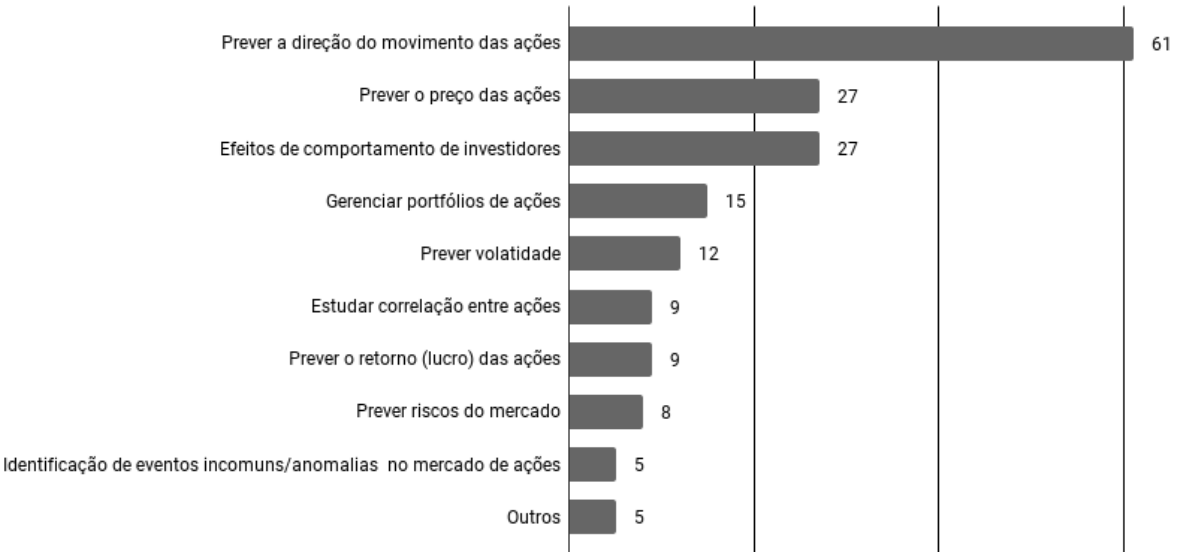


estudos produzidos foi alavancado, representando 55% das publicações (64 estudos). Em 2019, dos 25 estudos publicados, 12 foram em conferências e 13 em revistas, já em 2018, ano em que houve o maior número de publicações, os meios de publicações não foram tão equilibrados. A maioria (31 estudos) foi por meio de conferências, com apenas 8 em revistas.

QP2. Qual o propósito do estudo no mercado de ações?

A Figura 6 mostra que o principal objetivo dos estudos selecionados consiste em prever a direção do movimento das ações, concentrando aproximadamente 53% do total dos estudos. Os autores desenvolvem mecanismos para ajudar na tomada de decisão dos investidores, pois é possível identificar oportunidades de lucro ao vender ações em casos de futuras quedas de valor, comprar ações em casos de futuros aumentos ou ainda ou manter ações que possuem poucas alterações nos preços.

Figura 6 – Propósito dos Estudos no Mercado de Ações



Em (PICASSO et al., 2019), por exemplo, os autores combinam análise sentimental com predição de séries temporais utilizando técnicas de aprendizado de máquina a fim de prever

o movimento dos preços das ações de um portfólio. Os autores propuseram um modelo baseado em rede neural que reúne dados estatísticos e dados extraídos de notícias que podem afetar os preços de uma carteira formada por empresas listadas no índice NASDAQ. Para classificar o movimento dos valores, o modelo calcula as mudanças negativas e positivas de cada ação baseado nos resultados da análise de sentimentos aplicada nas notícias e nos fatores estatísticos identificados nos preços, representando possíveis lucros e prejuízos.

Outro exemplo é o estudo conduzido por Shah, Isah e Zulkernine (2019). Os autores utilizaram análise sentimental e criaram um dicionário de dados para comparar informações de notícias do site *MoneyControl*[2] que podem impactar nos preços das ações do índice *Nifty Pharma* e representa o setor farmacêutico do mercado financeiro da Índia. O modelo de análise de sentimentos de notícias baseado em dicionário classifica as ações em três categorias definidas de acordo com o movimento previsto dos preços, são elas: comprar, vender e manter. Essas classificações do movimento dos preços das ações ajudam o investidor nas suas escolhas, pois sinalizam antecipadamente o comportamento das ações no mercado financeiro.

Abrangendo aproximadamente 23% do total de artigos, outro propósito abordado pelos estudiosos foi a previsão do preço das ações, onde foram desenvolvidos mecanismos que são capazes de não apenas prever a direção do movimento, mas também o valor exato e/ou a faixa dos futuros preços. É o caso do estudo de Morris et al. (2019), onde é desenvolvido um algoritmo de aprendizagem de máquina de conjunto envolvendo *Long short-term memory* (LSTM), filtro de Kalman e regressão linear. Para validar a eficiência do algoritmo, foram utilizados dados dos preços finais de ações dos índices *Standard & Poor's 500 (S&P500)*, *Dow Jones Industrial Average (DJIA)* e do *Bitcoin (BTC)*, os resultados foram obtidos comparando os preços previstos pelo modelo e os valores reais das ações pro período estudado.

Também com 23% do total de artigos, o efeito do comportamento dos investidores no mercado de ações ficou entre os três propósitos mais estudados entre os estudos selecionados. Os pesquisadores procuram estudar o comportamento do investidor a fim de encontrar *insights* sobre o mercado acionário. Em (NI et al., 2019), por exemplo, os autores estudaram a relação entre o sentimento do investidor online e o desempenho do mercado de ações, a relação entre os preços das ações, o sentimento geral dos investidores e o efeito da avaliação de consultorias nos preços das ações e a opinião pública. Foi utilizada análise sentimental para extrair dados de textos dos investidores e realizadas análises de correlação. Os resultados mostraram que os preços sempre reagem ao sentimento em tempo real do investidor, portanto a previsão é mais eficiente quando feita com dados mais recentes e em curto prazo. A variável de avaliação de instituições de consultoria adicionada ao modelo mostrou que a opinião dos investidores não possuem relação com as avaliações, mas podem possuir relações com os preços das ações. Além disso, foi realizada uma análise de séries temporais que permitiu identificar que a opinião pública tem a capacidade de prever os preços das ações e é bem expressa em um período de duas semanas anteriores.

[2] *https://www.moneycontrol.com*

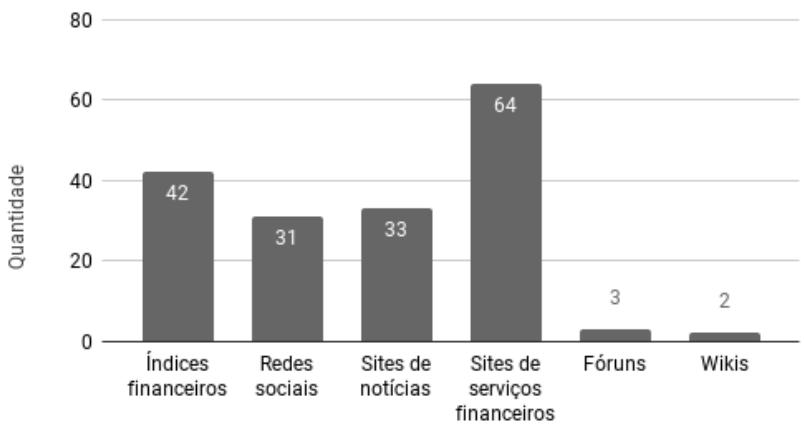
A gerência de portfólios de ações ficou com 13% dos estudos selecionados. Os pesquisadores propõem abordagens que consideram uma carteira de ações específica e buscam informações que possam indicar a decisão a ser tomada. Em sequência, o propósito de prever a volatilidade do mercado abordou 10% dos estudos. Liu (2019) realizou um experimento utilizando um portfólio de ações e técnicas de regressão para prever a volatilidade do mercado para as empresas listadas em um portfólio com o objetivo de maximizar os lucros dos investidores. A carteira escolhida no estudo abrange ações do índice S&P500. Foram extraídos dados históricos para treinamentos e validações dos modelos de aprendizagem profunda e de máquina utilizados, os modelos são: LSTM, GARCH e v-SVR. Os resultados mostraram que é possível prever a volatilidade da carteira com mais eficiência aplicando o aprendizado profundo usando LSTM RNNs e o aprendizado de máquina com v-SVR. O modelo GARCH obteve desempenho inferior em previsões de grandes intervalos de tempo.

Em relação a outros propósitos identificados neste mapeamento, estudar correlação entre ações ocupou 7% (9 estudos), prever o retorno (lucro) das ações 7% (9 estudos), prever riscos de mercado 6% (8 estudos), identificação de anomalias no mercado 4% (5 estudos) e outros 4% (5 estudos).

QP3. Quais são as fontes de dados utilizadas das quais os dados são coletados?

A QP3 tem como objetivo identificar quais são as fontes de dados utilizadas nos estudos selecionados. Também foi analisado se as fontes identificadas são internas ou externas. Como apresentado na Figura 7, os sites de serviços financeiros foram a fonte de dados mais utilizadas (55% - 64 estudos), seguida pelos índices financeiros (36% - 42 estudos), sites de notícias (28% - 33 estudos), redes sociais (26% - 31 estudos), fóruns (2% - 3 estudos) e wikis (1% - 2 estudos).

Figura 7 – Tipos de Fonte de Dados



Os sites de serviços financeiros proveem serviços relacionados ao mundo financeiro,

como API's que permitem acessar dados históricos de ações, preços de fechamento diários em um determinado período, relatórios, informações em tempo real sobre o movimento das ações, notícias e simulações. Dos 64 estudos que utilizam sites de serviços financeiros, o *Yahoo Finance* foi o site mais citado, aparecendo em 29 estudos, seguido pelo *Google Finance*, citado em 8 estudos.

Em (PENG, 2019), por exemplo, é proposta uma abordagem de aprendizado de máquina que aprende com dados anteriores para tomar decisões sobre quais ações são lucrativas para negociar. Para executar a abordagem, o autor utilizou dados históricos referentes a ações de petróleo na bolsa dos EUA. Os dados são os preços de fechamento diários da *United States Oil Fund* (USO) e foram obtidos em tempo real do *Yahoo Finance* no período de 2006 a 2019.

O *Google Finance* foi utilizado por Lee e Paik (2018) para recuperar dados de ações dos três principais índices financeiros da América: *Dow Jones*, NASDAQ e S&P500. Os autores criaram uma abordagem baseada em *Big Data* para prever a direção do movimento dos preços das ações baseado na correlação dos preços das ações, tweets e notícias. Os dados dos preços foram obtidos em tempo real através da API do *Google Finance*, as notícias foram obtidas dos sites da CNN[3], *New York Times*[4] e BCC[5] e os tweets foram coletados da API oficial do *Twitter*.

Os índices financeiros foram a segunda fonte de dados mais utilizadas entre os estudos selecionados. Os autores costumam utilizar dados históricos e/ou em tempo real de índices financeiros para realizar previsões futuras e comparar com os preços reais do mercado, a fim de provar a eficácia de suas abordagens. Em (GUO et al., 2017), por exemplo, os autores utilizam os dados históricos dos índice financeiros NASDAQ e *Shanghai Composite Index* para criar um modelo de previsão de séries temporais baseado nos algoritmos RBF e 2D LPP capaz de prever os preços das ações listadas nos índices.

Os sites de notícias foram a terceira fonte mais citada entre os autores. As notícias servem como um objeto de estudo para identificar correlações entre os acontecimentos do mundo real noticiados e os eventos no mercado de ações. Em (DHAS; VIGILA; STAR, 2018), os estudiosos combinaram análise técnica e sentimental para realizar previsões em tempo real do mercado de ações. Os dados da análise sentimental foram coletados do *Twitter* e de artigos de notícias publicados nos sites *Money Control*, *Financial Express* e *Economic Times*. Já os da análise técnica foram extraídos do histórico das ações da bolsa de valores da Índia, *National Stock Exchange of India* (NSE). Os resultados mostraram que as notícias e as mídias sociais interferem no movimento das ações e a análise de *Big Data* pode prever o mercado em tempo real.

O *Sina Weibo*, uma das maiores redes sociais chinesas, foi utilizada em (WANG; WANG, 2016) para extrair a opinião pública por meio da análise de sentimentos aplicada nos comentários sobre ações feitos por usuários da plataforma. A análise de sentimentos foi

[3] *https://cnn.com*
[4] *https://www.nytimes.com*
[5] *https://www.bbc.com*

realizada obtendo dados diários dos posts do *Sina Weibo*, do *Tong Hua Shun Network* e do *Dong Fang Cai Fu Network*, que são sites de serviços financeiros que também possuem uma plataforma de rede social para usuários registrados. Os autores criaram um vetor de palavras para identificar comentários que possuem relação com ações baseados em um dicionário de sentimentos e quantificaram o valor de cada emoção para as ações, obtendo uma previsão para o mercado de ações, os resultados mostram que a previsão dos preços das ações e os valores dos segmentos de emoção são muito sensíveis aos preços. As redes sociais foram identificadas como a quarta principal fonte de dados utilizada nos estudos selecionados.

De acordo com a Figura 7, os fóruns e as *wikis* foram as fontes de dados menos citadas pelos autores. Ma e Zhao (2017) usaram mineração de dados para coletar informações sobre a opinião pública relacionada com ações através de comentários em um fórum chinês, chamado *Xueqiu*[6]. Os estudiosos aplicaram análise sentimental baseada em palavras-chaves e quantificaram a ocorrência das palavras nos comentários coletados durante três dias. O foco do estudo foram as ações listadas no índice *SSE Composite Index*. Curmea et al. (2013) utilizaram os dados criados pelos usuários na Internet ao procurar por termos e tópicos de interesse. Os autores utilizam os dados do volume de pesquisa da *Wikipedia* e do *Google* a fim de encontrar relações com o movimento do mercado de ações. Eles descobriram que o aumento no volume de pesquisas por tópicos como política ou negócios precedem uma queda no mercado.

O assunto mais investigado foi a utilização dos dados históricos para previsões futuras no mercado acionário, com 81 estudos que utilizam grandes conjuntos de dados históricos de ações, representando cerca de 70% do total dos estudos selecionados. A aplicação de análise sentimental no domínio do mercado financeiro foi o segundo assunto mais investigado, sendo investigado em 40 estudos (aproximadamente 35%), sendo que a metade desses estudos (20 estudos) combinam dados históricos e dados obtidos por técnicas de análise sentimental. Os dados referentes a preços de ações e relatórios históricos são considerados dados internos, já os dados obtidos por análise sentimental, como os dados de notícias e mídias sociais são considerados dados externos.

(QP4) Quais técnicas, algoritmos, modelos e ferramentas foram empregados?

A questão de pesquisa 4 visa mapear as principais técnicas, algoritmos, modelos e ferramentas utilizados nos estudos selecionados. Para uma classificação mais completa, foi adicionado a este mapeamento os principais métodos e tecnologias identificados nos estudos.

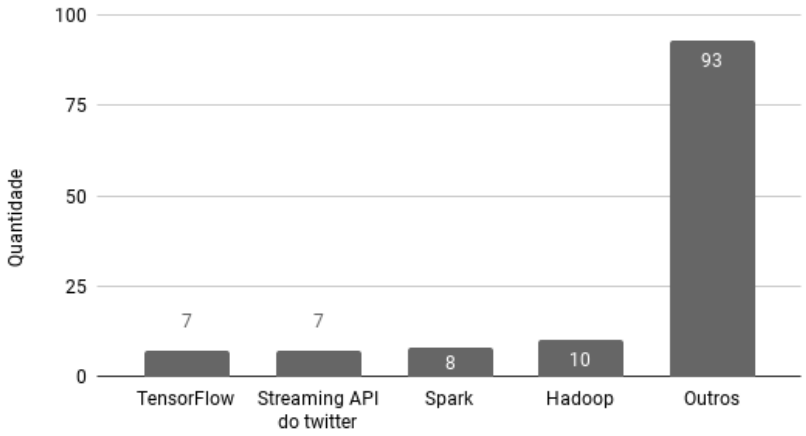
A Figura 8 apresenta as principais ferramentas adotadas nos estudos selecionados. As ferramentas podem ser API's, *frameworks*, programas e bibliotecas que permitem a realização de alguma tarefa do processo de *Big Data*, ou seja, as ferramentas são qualquer instrumento que auxilie na execução das etapas da abordagem. Foram encontradas 97 ferramentas ao

[6] *https://xueqiu.com*

todo, muitas são utilizadas em um único estudo e são de uso específico do estudo em que foram citadas, como dicionários, bibliotecas, API's, etc., o que sugere que existe uma grande variedade de ferramentas que podem ser exploradas nas abordagens de *Big Data* para o mercado financeiro. O *Hadoop*, por exemplo, foi o instrumento mais comumente utilizado entre os autores, sendo citado em 10 artigos (8%).

> O *Hadoop* é uma plataforma de *Big Data* de código aberto que fornece uma estrutura para lidar com grandes conjuntos de dados por meio de armazenamento e processamento distribuídos. A estrutura é baseada na suposição de que as falhas de hardware são comuns e, portanto, é projetada de forma a cuidar automaticamente de todas as possíveis falhas do sistema. (PENG, 2019, p. 309).

Figura 8 – Ferramentas Utilizadas nos Estudos Selecionados



O *Spark* foi a segunda ferramenta mais utilizada (7% - 8 estudos). Em Peng (2019), por exemplo, o *Spark* é utilizado para processar dados em tempo real de forma distribuída. A abordagem desenvolvida no estudo identifica ações com margem positiva de lucros. A abordagem foi desenvolvida no ecossistema *Hadoop* e o *Spark* foi utilizado no pré-processamento dos dados com a API do *Python*, chamada *PySpark*.

A API oficial do *Twitter* e o *TensorFlow* foram utilizados em 7 estudos cada um, representando 6% dos estudos. A API de *streaming* do *Twitter* permite que mensagens sejam coletadas em tempo real, com um provável atraso que deve ser considerado devido ao servidor (ROMANOWSKI; SKUZA, 2017). No estudo de Romanowski e Skuza (2017), os dados são obtidos do *Twitter* por meio da API de *streaming* em um período de três meses (janeiro e março de 2013). Os tweets coletados têm relação com as ações da APPLE, o objetivo é realizar a previsão dos movimentos dos preços das ações utilizando análise sentimental aplicada nos tweets coletados.

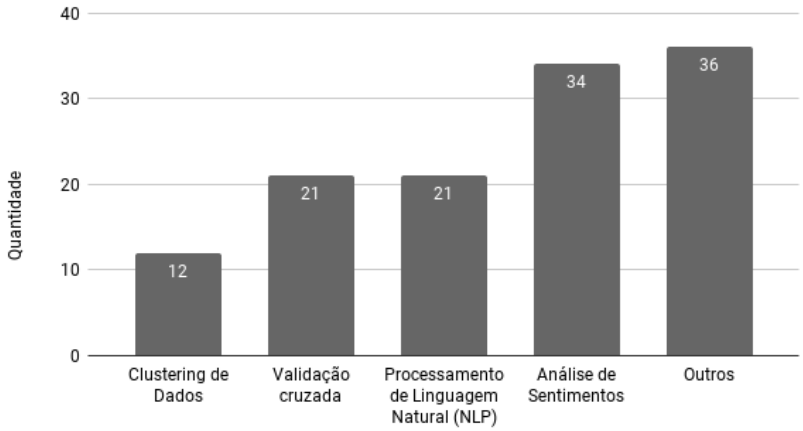
> O *TensorFlow* é uma biblioteca de software de código aberto para treinar máquinas em uma série de tarefas. É uma biblioteca simbólica da matemática e também é usada como um sistema para a construção e treinamento de redes neurais para a detecção e descriptografia de padrões e correlações,

> proporcionais ao aprendizado e raciocínio humanos. (MOURELATOS et al., 2018, p. 4).

Em (MOURELATOS et al., 2018), o *TensorFlow* é utilizado para implementar uma LSTM (*Long short-term memory*), que é uma rede neural de memória de longo prazo capaz de prever e modelar índices financeiros. O experimento foi feito utilizando dados do índice *FTSE/Athex Large Cap*, além de uma comparação com uma abordagem híbrida que utiliza algoritmos genéticos e máquinas de vetor de suporte (GA-SVR).

A Figura 9 apresenta as principais técnicas encontradas nos estudos selecionados. A principal técnica utilizada foi a análise de sentimentos, com 34 estudos (aproximadamente 29%). A análise sentimental é uma técnica poderosa para obter opiniões e rastrear atitudes de usuários a partir de dados gerados por diferentes meios de comunicação e expressão na web, como avaliações, comentários, bate-papos, compartilhamentos e tweets (CHEN; ZHENG, 2018).

Figura 9 – Técnicas Utilizadas nos Estudos Selecionados



Em (CHEN; ZHENG, 2018), a análise sentimental foi aplicada em dados obtidos em tempo real da API de streaming do *Twitter* com o objetivo de extrair informações que possam classificar os tweets relacionados a uma companhia em três categorias: positivo, negativo ou neutro. Foi desenvolvido um analisador de sentimentos e implementado um modelo *Multilayer Perceptron* (MLP) para prever o preço futuros das ações de uma determinada empresa baseado na opinião pública e nos preços históricos. Os resultados fornecem análises descritivas e preditivas simultaneamente, de modo que os investidores podem utilizar as previsões de tendências e enriquecer as decisões com *insights* obtidos em tempo real para uma escolha mais bem informada.

A segunda técnica mais utilizada foi o Processamento de Linguagem Natural (PLN), mencionado em 21 estudos (18%). A técnica PLN é uma sub-técnica da análise de sentimentos utilizada na fase de processamento dos dados. O PLN é o processo de análise e interpretação de texto:

> A análise de sentimentos geralmente se refere ao processamento de linguagem natural em termos de análise e interpretação de texto e linguística computacional para extrair, identificar, caracterizar o sentimento (um ponto de vista associado ou atitude em relação) de um determinado texto. (ROMANOWSKI; SKUZA, 2017, p. 11).

Li et al. (2017) utilizou o PLN pra extrair informações de tweets e classificar a opinião em positiva, neutra ou negativa a fim de prever o movimento das ações listadas no índice NASDAQ. Nessa abordagem, foi proposta uma abordagem baseada em *bag-of-words*.

Também citado por 21 estudos, a validação cruzada é umas das principais técnicas utilizada para validar modelos de aprendizado de máquina. Em (SHAH; ISAH; ZULKERNINE, 2019), por exemplo, para prever o movimento das ações do setor farmacêutico com base em artigos de notícias, os autores desenvolveram uma abordagem baseada em dicionário e usaram validação cruzada para medir a eficiência do modelo em relacionar os sentimentos obtidos das notícias com os preços das ações.

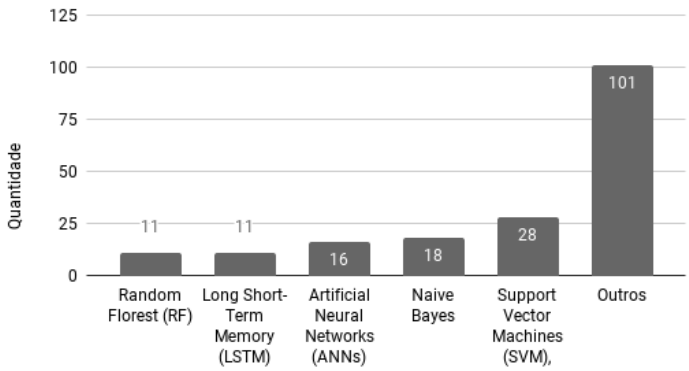
O clustering de dados foi explicitamente utilizado em 12 estudos, representando aproximadamente 10% do total. Essa técnica trata-se do processo de organizar dados que são semelhantes ou que atendam aos mesmos critérios de acordo com alguma métrica de similaridade (CAO et al., 2018). Utilizando *Hadoop*, uma abordagem de mineração de dados foi proposta em Cao et al. (2018) implementando a técnica de clusterização para agrupar informações obtidas através de um *web crawler* a fim de descobrir a real intenção de negociação dos investidores do mercado chinês.

A Figura 10 apresenta os principais modelos e algoritmos identificados nos estudos. Os modelos e algoritmos são apresentados em uma única classificação devida a dificuldade de diferenciar os termos, pois em muitos casos os autores utilizam tanto o termo algoritmo quanto o termo modelo para um mesmo item, já que um modelo é implementado através de um algoritmo. Dessa forma, o autor pode se referir ao SVM como algoritmo em um momento e posteriormente como um modelo.

O *Support Vector Machine* (SVM) foi o algoritmo mais utilizado entre os estudos selecionados, mencionado em 28 trabalhos (24%). O SVM é um algoritmo de aprendizado de máquina supervisionado que parte do princípio VC-dimensional da teoria de aprendizagem estatística e do risco estrutural mínimo. Esse algoritmo pode encontrar o melhor cenário entre a complexidade das amostras (precisão de aprendizagem) e a capacidade de aprender (identificar qualquer amostra sem erros). O SVM tem um papel importante quando se trata de estimativas de regressão de função, reconhecimento de classificação de padrão, tendência de estoques e sequência de tempo de previsão (YANG et al., 2020).

No estudo de Yang et al. (2020), o SVM foi utilizado para compor uma abordagem de previsão de volatilidade do mercado de ações. A intenção é identificar possíveis riscos de mercado com previsões de curto prazo. A abordagem foi aplicada no mercado financeiro da China com os dados das flutuações dos preços das ações de empresas listadas no índice CSI 300. Os resultados mostraram que o SVM tem um bom efeito de predição.

Figura 10 – Modelos e Algoritmos



O classificador *Naive Bayes* foi o segundo algoritmo mais utilizado entre os estudos mapeados, aparecendo em 18 trabalhos (15%). Em (SEIF; HAMED; HEGAZY, 2018), o algoritmo é utilizado na criação de um modelo capaz de prever tendências de ações baseado em notícias, tweets e preços históricos. Para analisar as informações de tweets e notícias, foi realizada uma análise sentimental com base no classificador bayesiano.

> *Naive Bayes* é uma técnica de classificação baseada no Teorema de Bayes com uma suposição de independência entre os preditores. Em termos simples, um classificador *Naive Bayes* assume que a presença de um determinado recurso em uma classe não está relacionada à presença de qualquer outro recurso. O modelo Naïve Bayes é fácil de construir e particularmente útil para conjuntos de dados muito grandes. Junto com sua simplicidade, *Naive Bayes* é conhecido por ter um desempenho superior, mesmo com métodos de classificação altamente sofisticados. (SEIF; HAMED; HEGAZY, 2018, p. 677).

Alguns autores citam o conjunto de modelos de redes neurais artificiais, muitos realizam comparações entre mais de um tipo de rede neural, assim, esse conjunto de modelos computacionais foi citado em 16 estudos selecionados, representando 14% do total de estudos. No estudo de Assis, Pereira e Silva (2018), por exemplo, os autores utilizam três modelos de redes neurais: *feedforward, cascade forward* e *Elman*. As redes neurais artificiais são baseadas em neurônios biológicos:

> A tendência recente na análise de *Big Data* é criar um tipo de cérebro humano virtual de uma configuração que pode pegar os dados como entrada e processá-los com um pouco de suporte para ajudar a organização a fazer análises automatizadas ou semiautomáticas resultando em decisões quase em tempo real. (VENKADACHALAM et al., 2018, p. 3).

O *Long Short-Term Memory* (LSTM) e o *Random Florest* (RF) também foram comumente utilizados nos estudos. Cada algoritmo foi mencionado em 11 estudos, ou seja, está presente em 9% do total de estudos. O *Random Forest* é um algoritmo de classificação. Yang, Liu e Wu (2018) aplicaram o RF para criar uma abordagem de previsão de retornos futuros de ações com base em fatores de ganho encontrados em preços históricos classificados a partir do algoritmo, o modelo desenvolvido recomenda ações do índice S&P500 dinamicamente.

O LSTM é um modelo de rede neural artificial baseado em memória de longo prazo, um exemplo de sua aplicação pode ser observado no estudo feito por Sismanoglu et al. (2019), onde a rede é aplicada em ações do NYSE e do NASDAQ. Segundo os autores, nas redes LSTM, cada neurônio possui sua própria memória, de modo que os dados passados possam ser armazenados e utilizados no processo de aprendizagem do modelo. O modelo de *bag-of-words* também foi mencionado em 11 estudos. O *bag-of-words* se trata de um classificador utilizado em análise de sentimentos de textos. Em (ROMANOWSKI; SKUZA, 2017), por exemplo, os autores utilizam o modelo para classificar os tweets coletados pela API de streaming do *Twitter* na etapa de análise sentimental após o pré-processamento dos dados.

Os métodos mais utilizados entre os 115 estudos foram o *Principal Component Analysis* (PCA) (8% - 10 estudos) e o *Time Slice Window* (7% - 9 estudos). O PCA trata se de um método para redução de dimensionalidade de grandes conjuntos de dados que extrai e mantém apenas as informações mais relevantes, sem perder o sentido original dos dados (WENG et al., 2018). O *Time Slice Window* é amplamente utilizado em problemas de reconhecimento de padrões, esse método permite que a leitura dos dados de uma transação sejam segmentados em diferentes janelas de tempo (CAO et al., 2018).

Em relação as tecnologias, os resultados mostraram que as mais utilizadas são as linguagens de programação *Python* e R, mencionadas em 20 e 16 (17% e 14%) estudos respectivamente. O *Python* acaba sendo muito utilizado, pois oferece muitas ferramentas e recursos para aprendizado de máquina (COYNE; MADIRAJU; COELHO, 2017). O R é um ambiente e uma linguagem de programação adequado para várias análises de dados. Essa linguagem é capaz de resolver problemas estatísticos com tarefas simples (TSENG et al., 2018).

(QP5) Quais das etapas do Big Data são contempladas no artigo?

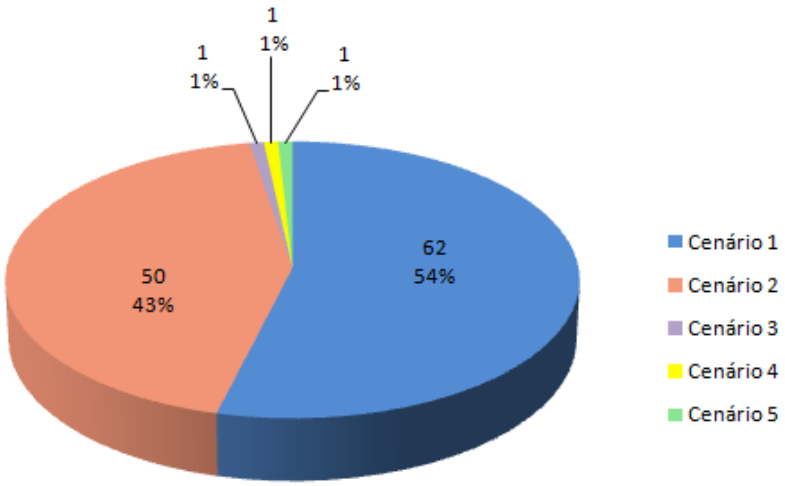
Cada um dos 115 artigos selecionados foram analisados de forma a identificar as etapas de *Big Data* que foram contempladas no estudo. Ao todo foram identificados 5 diferentes combinações das etapas do *Big Data* utilizadas nos estudos. Essas combinações foram chamadas de cenários. A Tabela 2 apresenta os cenários que foram identificados. As etapas marcadas com "X" foram contempladas no estudo, já as marcadas com "-" se tratam de etapas não contempladas.

Tabela 2 – Descrição dos Cenários.

Cenário	Coleta	Armazenamento	Processamento	Análise	Visualização
Cenário 1	X	X	X	X	X
Cenário 2	X	X	X	X	-
Cenário 3	-	X	X	X	X
Cenário 4	-	-	X	X	-
Cenário 5	X	-	-	X	-

A Figura 11 apresenta a quantidade e a porcentagem de artigos para cada cenário identificado. O cenário mais recorrente foi o 1, em que todas as fases são abordadas. O segundo cenário mais identificado foi o 2, onde a visualização não é contemplada. Já os cenários 3, 4 e 5 foram os menos identificados, com ocorrência em um único artigo cada um.

Figura 11 – Quantidade de Estudos por Cenário



A seguir são apresentados alguns estudos exemplificando cada um dos cenários identificados.

No cenário 1, todas as etapas são contempladas. Os autores realizam o processo de *Big Data* desde a coleta dos dados até a visualização. É o caso do trabalho de Martinez, roman e Casado (2018), onde os autores desenvolveram um sistema de *Big Data* para ordens de negociações levando em conta o sentimento dos investidores. Os dados foram coletados pela plataforma *Apara*[7] de diferentes fontes: televisão, rádio, notícias e redes sociais. Os dados foram armazenados e técnicas de PLN foram utilizadas com classificadores de probabilidade guiados por aprendizado supervisionado. O retorno compreende três categorias: positiva, negativa ou neutra. Os textos de treinamento são tratados com PLN para extrair as características mais importantes. Os autores usam redes bayesianas para implementar o modelo na plataforma do *InvestMood*. Os algoritmos utilizados são fornecidos na plataforma *dVelox*. Os resultados das previsões são mostrados em gráficos que mostram as informações dos preços das ações relacionadas com o humor dos investidores.

Outro exemplo é o estudo de Day, Cheng e Li (2018), onde é desenvolvida uma abordagem de robô advisor para serviços financeiros com *Big Data*. Os autores criam um módulo de otimização de portfólio baseado em diferentes fontes, como preços de ações e perfil de investidor. Os autores detalham cada etapa da análise de *Big Data*. Na coleta, os dados de fechamento diário ajustados são obtidos automaticamente e o armazenamento é feito num

[7] *http://www.apara.es/*

arquivo com uma numeração de série exclusiva para ser acessado através de um banco de dados. Os dados são pré-processados e transformados de séries temporais para matrizes capazes de serem processadas por um módulo de LSTM. Em seguida, o modelo é treinado processando os dados de entrada (etapa de processamento) e comparando suas previsões com os valores verdadeiros fornecidos (etapa de análise). A saída é o modelo treinado com um histórico de treinamento. A visualização é fornecida através de figuras de flutuações de retorno de logaritmo e movimentos de preços, matriz de confusão para carteiras e componentes de ETF e retornos cumulativos.

Attigeri et al. (2015) também desenvolveram um estudo que aborda todas as fases do processo de *Big Data*. Os autores desenvolveram uma abordagem para previsão de tendências de ações utilizando dados do *Twitter*, de notícias e de preços históricos. O modelo foi implementado na plataforma do *Hadoop* para ações de uma determinada empresa. Os dados de notícias foram coletados utilizando um rastreador de web, o *Mozenda Web Crawler*, os do *Twitter* foram obtidos pela API que a rede social disponibiliza e os dados históricos foram fornecidos pelo *Yahoo Finance*. No total foram armazenados 355 artigos de notícias e 430 tweets. Os autores utilizaram dados de uma semana, a previsão é em relação ao dia posterior. Os dados foram preparados com processo de lemmatização, que tem o objetivo de reduzir formas de dados diferentes a uma forma comum, remoção de palavras comuns que não expressam sentimentos, remoção de URL's e remoção de duplicatas. Com os dados preparados, um algoritmo foi executado para relacionar as palavras encontradas à palavras de um dicionário base, a fim de classificar os sentimentos. Em seguida, os sentimentos de tweets e notícias são agregados a fim de gerar o sentimento para a companhia. Por fim, os resultados são visualizados através da união do *Hadoop* com o R, que produzem gráficos que exprimem a relação dos sentimentos obtidos com as tendências das ações. Os dados gerados também são comparados com dados reais dos preços das ações da empresa.

O cenário 2 é composto por todas as fases, exceto a visualização. Muitos autores desenvolvem uma abordagem, discutem, executam e analisam os resultados, mas não desenvolvem uma forma de visualização computacional, alguns apresentam os resultados nos artigos em forma de tabela, sem uma representação gráfica. Em Shih et al. (2016), por exemplo, os autores desenvolvem um modelo de tendências de ações baseado em redes de Petri, O objetivo é identificar sinais de compra, venda e neutralidade nas ações. Os dados utilizados são informações históricas de preços de ações e indicadores obtidos dos sites do *Capital Finance*[8] e do Masterlink[9], os autores escolhem 14 indicadores técnicos e os armazena para entrada do modelo. O processamento dos dados é realizado usando o PCA para dividir o conjunto e redimensionar, assim são utilizados apenas os componentes principais para explicar a quantidade máxima de variância, após aplicar o PCA, são construídos 3 conjuntos de dados, o primeiro com três indicadores, o segundo com 7 e o último com 2. A análise é realizada com o modelo

[8] *www.capital.com.tw*
[9] *www.masterlink.com.tw*

construído de *Rough Petri Nets* (RPN). São definidas estratégias para identificar os sinais e produzir a saída do modelo, que pode ser compra, venda ou segura. Após a análise, os autores não trouxeram nenhuma solução gráfica para apresentar os resultados, no artigo os dados foram mostrados em tabelas.

Outro exemplo em que ocorre o cenário 2 é no estudo (SEZER; OZBAYOGLU; DOGDU, 2017), onde é proposto uma abordagem de rede neural com aprendizado supervisionado baseado em algoritmo MLP para prever sinais de compra e venda de ações utilizando indicadores técnicos e preços históricos. Os autores utilizam o *Spark* para desenvolver o modelo, os dados de ações do índice DJIA são coletados do *Yahoo Finance*. Os dados armazenados são normalizados e com base no preço de fechamento diário. O sinal de compra ou venda de cada dia do conjunto histórico é definido para treinar o modelo, identificando as suas categorias. A análise é feita considerando os indicadores técnicos, que são calculados pelo modelo para cada ação, a saída consiste nos sinais de compra ou venda. Os autores exemplificam a utilização da abordagem e disponibilizam uma tabela contendo as informações da previsão, mas não chegam a desenvolver uma representação para visualizar os resultados da abordagem.

A abordagem de Weihua e Qin (2018) também é um exemplo do cenário 2. Os autores criaram um método de análise de preço de ações e circunstância de negócios para uma determinada empresa. A empresa escolhida pelos pesquisadores foi a *CityMedia*, listada na Bolsa de Valores de Xangai. O *dataset* é formado por registros de negociação de ações da empresa e informações financeiras coletados do *Eastmony* e do *CITIC Securities*. Os dados são extraídos e armazenados criando páginas de índices de frases que contém as palavras chaves relacionadas à ações. Os dados são pré-processados utilizando *JavaScript* (JS) no navegador, a página HTML é lida e arquivada em arquivos XHTML. Na etapa de processamento, os dados são minerados utilizando a técnica de frequência de palavras, o conteúdo do texto é identificado e classificado de acordo com o tipo de informação. A análise foi feita com regressão múltipla utilizando os dados de preços de dias anteriores, de informações da empresa, de transações, entre outros. O resultado da previsão é apresentado ao leitor em uma tabela, os autores não desenvolveram uma solução capaz de apresentar os resultados.

No cenário 3, todas as etapas são contempladas, exceto a coleta dos dados. O trabalho que se encaixa nesse caso é feito por Venkadachalam et al. (2018), onde os autores criam um modelo para prever os retornos das ações intradiários. Os dados utilizados nesse estudo são armazenados em arquivos .csv, mas os autores não explicam de onde eles são obtidos, apenas que se tratam de dados de retornos de ações, portanto, a primeira etapa abordada é o armazenamento dos dados. Os dados são preparados excluindo valores ausentes, identificando variáveis e normalizando os conjuntos. Em seguida, na fase de processamento, é proposto um modelo de rede neural baseado em MLP, um aprendizado de máquina supervisionado. Para treinar o modelo os dados são divididos em 20% para conjunto de treino e 80% para conjunto de validação usando o *Pareto Principle*, o algoritmo é implementado usando o R. A análise dos dados é feita atribuindo pesos as variáveis identificadas ao prever uma nova variável dependente.

A etapa de visualização é contemplada com a plotagem de gráficos que ilustram os resultados das previsões. Para validação, os autores comparam os preços previstos com os preços atuais das ações.

O cenário 4 só contempla as fases de processamento e análise dos dados, ou seja, os autores não especificam de onde os dados são coletados nem como são armazenados, tampouco a visualização dos resultados. Tal cenário ocorreu somente no estudo de Tang e Fan (2016). Os autores desenvolveram um modelo baseado em rede neural e algoritmo genético para prever o preço de uma ação individualmente. No estudo, a fase de coleta e armazenamento dos dados não são abordadas. É mencionado no estudo que os dados são de preços históricos e que são processados realizando uma separação cronológica dos dados e calculando a taxa de erro e caso ocorra um crescimento da taxa de erro, significa que existe uma mudança na estrutura do mercado. A análise proposta é feita utilizando a taxa de erro e os pesos dos nós do algoritmo genético produzido pelo resultado da rede neural. A visualização não é abordada, os autores explicam que o artigo apenas apresenta o modelo, mas não aplica com dados do mundo real.

O último cenário possui as etapas de coleta e análise. Os autores propõem uma abordagem utilizando grandes conjuntos de dados do mercado de ações mas não apresentam o ponto de vista computacional das etapas de armazenamento, processamento e visualização. Esse caso ocorre no trabalho de Yoon, Suge e Takahashi (2018), em que a influência dos artigos de notícias é analisada no mercado de ações da Coreia. Os dados foram obtidos da plataforma do *Thomson Reuters*. Foram coletados mais de 34 mil linhas de dados de séries temporais contendo informações como volume de negociações e preços de compra e venda de ações de cinco empresas (*Samsung Electronics, Hyundai Motors, POSCO, Hyundai Mobis e Kia Motors*). Os artigos de notícias foram obtidos pelo nome das empresas ou pelos símbolos de cotação e compreendem sete idiomas diferentes. As ocorrências de notícias são distribuídas em duas distribuições temporais, sendo uma ao longo de 24 horas e uma no intervalo de uma hora. A análise também é realizada calculando o impacto do volume de negociações para o mesmo período de tempo das notícias. Para visualizar os resultados, os autores usaram casos de estudo do mundo real para prever a volatilidade dos preços. Os resultados foram apresentados em gráficos que mostram as quedas e aumentos dos preços.

5 CONCLUSÃO

Este capítulo apresenta as principais considerações, contribuições, bem como as limitações encontradas durante a pesquisa. Ao final, também são apresentadas algumas sugestões para realização de trabalhos futuros.

5.1 CONSIDERAÇÕES GERAIS

O objetivo geral desse trabalho foi sumarizar as principais abordagens computacionais para *Big Data* aplicadas no mercado de ações a fim de identificar o estado da arte na área.

Considerando os resultados alcançados até o momento a partir das atividades conduzidas, é possível afirmar que o campo de pesquisa para aplicação de *Big Data* é um tópico recente, com maior concentração de estudos nos dois últimos anos compreendidos pela pesquisa (2018 e 2019). Além disso, os artigos de conferência são os principais meios de publicação. Para os pesquisadores, prever preços e/ou movimentos de ações é o principal propósito dos estudos, para isto, a influência do comportamento e a opinião pública dos investidores se torna um dos principais assuntos, além da gerência de carteiras de investimento que também é bastante investigada. Para a condução das abordagens, os sites de serviços financeiros e os índices financeiros são as principais fontes de dados históricos, já os sites de notícias e as redes sociais são as principais fontes de dados em tempo real e com maior variedade de informações, muitos estudos combinam dados históricos e dados de notícias e/ou redes sociais em suas abordagens.

A partir dos resultados dos estudos selecionados, é possível observar que os pesquisadores estão investindo na criação de novas ferramentas (97 ferramentas). No entanto, muitas ferramentas foram utilizadas em um único estudo. As ferramentas mais comumente utilizadas são a plataforma *Hadoop*, o *framework Spark*, a biblioteca *TensorFlow* e a API do *Twitter*. Em relação às técnicas utilizadas, análise de sentimentos e o processamento de linguagem natural são as principais técnicas para extrair informações valiosas da web. A validação cruzada é utilizada com frequência para avaliar os modelos desenvolvidos. Para lidar com o grande volume de dados, muitos autores utilizam a técnica de clusterização dos dados. Diversos modelos de redes neurais são amplamente utilizados nas abordagens. Já em relação aos algoritmos, o SVM possui alta capacidade de performar classificações de tendências e padrões, por isso é o algoritmo mais aplicado. O algoritmo de *Naive Bayes* também é bastante utilizado, pois é um classificador simples que consegue lidar com grandes conjuntos de dados mantendo um bom desempenho. A fim de reduzir a dimensão dos grandes conjuntos de dados, o PCA é um método bastante eficaz pois é capaz de identificar apenas as informações mais relevantes presentes nos dados e extraí-los sem perder valores importantes. No reconhecimento de padrões, o *Time Slice Window* é bastante utilizado pois consegue diminuir a complexidade do volume dos dados realizando leituras em diferentes janelas de tempo. Em relação às tecnologias utilizadas, é

possível afirmar que o *Python* e o R são as mais utilizadas para aprendizados de máquina e análise de dados com problemas estatísticos, que são frequentes na análise do mercado de ações.

Por fim, a respeito das etapas do processo de *Big Data*, conclui-se que a etapa de visualização é a mais difícil de ser abordada e implementada. Muitos autores realizam as etapas anteriores a visualização, mas não desenvolvem uma apresentação computacional que sumarize os resultados, apenas os discutem. Não houveram ocorrências significativas da ausência de outras etapas.

5.2 PRINCIPAIS CONTRIBUIÇÕES

Este trabalho de conclusão de curso possibilitou sumarizar as principais evidências associadas à abordagens computacionais para *Big Data* aplicadas no mercado de ações. Além disso, o mapeamento sistemático fez parte de um projeto de iniciação científica o qual permitiu a elaboração e publicação do seguinte artigo:

> MACEDO, N. L. M.; SOUZA, E. F.; MEINERZ, G. V. (2020). Abordagens computacionais para Big Data aplicadas ao mercado financeiro: Um mapeamento sistemático. **XXIV Seminário de Iniciação Científica e Tecnológica da UTFPR**.

5.3 LIMITAÇÕES DA PESQUISA

Algumas das principais limitações e dificuldades encontradas neste trabalho são listadas a seguir:

- No mapeamento sistemático foi utilizado apenas o banco de dados da Scopus para busca dos estudos. Embora a Scopus seja considerada a maior base de dados de resumos e citações, é possível que alguns estudos valiosos tenham ficado de fora da análise. Mesmo assim, acredita-se que os estudos analisados no mapeamento fornecem uma visão geral sobre as pesquisas existentes relacionadas as abordagens computacionais para *Big Data* aplicadas no mercado de ações.
- O pouco conhecimento da aluna deste trabalho na área de *Big Data* e mercado de ações tornou a fase de análise mais complexa. Dessa forma, o tempo gasto na condução do mapeamento foi maior que o planejado.
- O período da pesquisa compreende os estudos publicados até 2019, ano em que foi iniciado este trabalho. Dado a limitação relacionada ao conhecimento da área, o envolvimento da aluna com o estágio, mudança de cidade e final de curso de graduação, não foi possível atualizar o período para os anos de 2020 e 2021.

5.4 TRABALHOS FUTUROS

Para a continuidade deste trabalho de pesquisa, considera-se utilizar outras bases de dados no processo do mapeamento sistemático da literatura. Com isso outros estudos relevantes podem ser retornados diminuindo as chances de trabalhos importantes ficarem de fora da análise. Entrevistas com especialistas da área confrontando os resultados do estudo, podem permitir uma visão mais ampla dos principais achados. Além disso, pretende-se aprofundar a especificidade das fontes de dados coletadas pelos artigos selecionados, fornecendo uma visão mais detalhada do negócio, a fim de identificar quais são os critérios para os dados serem considerando pelos estudos.

Referências

AKOKA, J.; COMYN-WATTIAU, I.; LAOUFI, N. Research on big data - a systematic mapping study. **Computer Standards & Interfaces**, v. 54, p. 105–115, 2017. Citado 3 vezes nas páginas 1, 13 e 14.

ASSIS, J. de M.; PEREIRA, A. C. M.; SILVA, R. C. e. Designing financial strategies based on artificial neural networks ensembles for stock markets. In: **2018 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2018. p. 1–8. Citado na página 28.

ATTIGERI, G. et al. Stock market prediction: A big data approach. In: . [S.l.: s.n.], 2015. p. 1–5. Citado na página 31.

AURUM, A.; DANESHGAR, F.; WARD, J. Investigating knowledge management practices in software development organisations - an australian experience. **Information and Software Technology**, v. 50, p. 511–533, 2008. Citado na página 4.

BACH, M. P. et al. Text mining for big data analysis in financial sector: A literature review. **Sustainability**, p. 27, 2019. Citado na página 14.

BATISTA, F. F. **O Desafio da Gestão do Conhecimento nas áreas de Administração e Planejamento das Instituições Federais de Ensino Superior**. Brasília: [s.n.], 2006. Citado na página 4.

BERNARDO, I.; HENRIQUES, R.; LOBO, V. Social market: Stock market and twitter correlation. In: SPRINGER. **International Conference on Intelligent Decision Technologies**. Vilamoura, Portugal: Springer, 2017. p. 341–356. Citado na página 7.

BEYER, M.; LANEY, D. The importance of big data: A definition. In: GARTNER. [S.l.], 2012. p. 7. Citado na página 5.

BHARDWAJ, A.; SINGH, W. Systematic review of big data analytics in governance. In: INTERNATIONAL CONFERENCE ON INTELLIGENT SUSTAINABLE SYSTEMS (ICISS). [S.l.], 2017. p. 501–506. Citado na página 14.

BHAVSAR, P. et al. Machine learning in transportation data analytics. In: _____. [S.l.: s.n.], 2017. p. 283–307. ISBN 9780128097151. Citado na página 8.

BUKOWITZ, W.; WILLIAMS, R. L. **The knowledge management fieldbook**. Great Britain: Financial Times Prentice Hall, 2000. Citado na página 4.

CAO, S.-I. et al. A stock trading intention recognition model based on data clustering. In: . [S.l.: s.n.], 2018. p. 338–343. Citado 2 vezes nas páginas 27 e 29.

CHEN, C.-I. P.; ZHENG, J. Improved big data analytics solution using deep learning model and real-time sentiment data analysis approach. In: **BICS**. [S.l.: s.n.], 2018. Citado na página 26.

CHEN, C. P.; ZHANG, C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on big data. **Information Sciences**, v. 275, p. 314–347, 2014. Citado na página 7.

COYNE, S.; MADIRAJU, P.; COELHO, J. Forecasting stock prices using social media analysis. In: . [S.l.: s.n.], 2017. p. 1031–1038. Citado na página 29.

CURMEA, C. et al. Quantifying the semantics of search behavior before stock market moves. **Proceedings of the National Academy of Sciences of the United States of America**, v. 111, p. 11600–11605, 8 2013. Citado na página 24.

DALKIR, K. **Knowledge Management in Theory and Practice**. Burlington, MA: Elsevier, 2005. Citado na página 4.

DAVENPORT, T. H.; PRUSAK, L. **Working knowledge**: how organizations manage what they know. 2. ed. Boston, USA: Harward Business School Press, 2000. Citado 3 vezes nas páginas 1, 3 e 4.

DAY, M.-Y.; CHENG, T.-K.; LI, J.-G. Ai robo-advisor with big data analytics for financial services. In: . [S.l.: s.n.], 2018. p. 1027–1031. Citado na página 30.

DHAS, J. L. J.; VIGILA, S. M. C.; STAR, C. E. Forecasting of stock market by combining machine learning and big data analytics. **Communications in Computer and Information Science**, p. 385–395, 2018. Citado na página 23.

EDUCATION SERVICES EMC. **Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data**. Wiley, 2015. ISBN 9781118876053. Disponível em: <https://books.google.com.br/books?id=J94WBgAAQBAJ>. Citado na página 6.

ESTEVES, S. R. M. **Requisitos de software funcionais para o desenvolvimento de plataforma digital de diagóstigo da gestão do conhecimento nas organizações**. Dissertação (Masters Dissertation (In Portuguese)) — UniCesumar, Maringá, Paraná, Brazil, 2017. Citado na página 4.

FORTUNA, E. **Mercado Financeiro**: Produtos e serviços. 11. ed. Rio de Janeiro, Brasil: Qualitymark Editora, 1998. Citado 2 vezes nas páginas 10 e 11.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.: s.n.], 2016. url http://www.deeplearningbook.org. Citado na página 8.

GUO, Z. et al. Financial index time series prediction based on bidirectional two dimensional locality preserving projection. In: . [S.l.: s.n.], 2017. p. 934–938. Citado na página 23.

KITCHENHAM, B. Procedures for performing systematic reviews. **Keele, UK**, v. 33, n. TR/SE-0401, p. 28, 2004. Citado na página 12.

KITCHENHAM, B.; CHARTERS, S. **Guidelines for performing Systematic Literature Reviews in Software Engineering**. Durham, UK, 2007. v. 2, n. EBSE 2007-001, 1051 p. Citado 2 vezes nas páginas 2 e 11.

LAIGNER, R. N. et al. A systematic mapping of software engineering approaches to develop big data systems. In: EUROMICRO CONFERENCE ON SOFTWARE ENGINEERING AND ADVANCED APPLICATIONS. Brazil, 2018. p. 446–453. Citado na página 14.

LEE, C.; PAIK, I. Stock market analysis from twitter and news based on streaming big data infrastructures. In: . [S.l.: s.n.], 2018. p. 312–317. Citado na página 23.

L'HEUREUX, A. et al. Machine learning with big data: Challenges and approaches. **IEEE Access**, v. 5, p. 7776–7797, 2017. Citado na página 5.

LI, B. et al. Discovering public sentiment in social media for predicting stock movement of publicly listed companies. **Information Systems**, PERGAMON-ELSEVIER SCIENCE LTD, v. 69, p. 81–92, set. 2017. ISSN 0306-4379. Citado na página 27.

LIU, Y. Novel volatility forecasting using deep learning–long short term memory recurrent neural networks. **Expert Systems With Applications**, v. 132, p. 99–109, apr 2019. Citado na página 22.

MA, R.; ZHAO, H. Predicting the change of stock market index based on social media analysis. In: . [S.l.]: Springer International, 2017. p. 154–162. Citado na página 24.

MARQUESONE, R. **Big Data: Técnicas e tecnologias para extração de valor dos dados**. Brasil: Casa do Código, 2016. ISBN 9788555192326. Disponível em: <https://books.google.com.br/books?id=cbWIDQAAQBAJ>. Citado 5 vezes nas páginas 1, 5, 6, 7 e 8.

MARSLAND, S. **Aprendizado de máquina: uma perspectiva algorítmica**. 2nd. ed. [S.l.: s.n.], 2014. ISBN 1466583282, 9781466583283. Citado na página 8.

MARTINEZ, R. G.; ROMAN, M. prado; CASADO, P. Big data algorithmic trading systems based on investors' mood. **Journal of Behavioral Finance**, v. 20, p. 1–12, 12 2018. Citado na página 30.

MORRIS, K. J. et al. Token-based adaptive time-series prediction by ensembling linear and non-linear estimators: A machine learning approach for predictive analytics on big stock data. In: . [S.l.: s.n.], 2019. p. 1486–1491. Citado na página 21.

MOURELATOS, M. et al. Financial indices modelling and trading utilizing deep learning techniques: The athens se ftse/ase large cap use case. In: **2018 Innovations in Intelligent Systems and Applications (INISTA)**. [S.l.: s.n.], 2018. p. 1–7. Citado na página 26.

NAKAGAWA, E. et al. **Revisão Sistemática da Literatura em Engenharia de Software: Teoria e Prática**. Brazil: Elsevier Brasil, 2017. ISBN 9788535285970. Citado 2 vezes nas páginas 11 e 12.

NETO, A. A. **Mercado Financeiro**. 11. ed. São Paulo, Brasil: Editora Atalas S.A., 2012. ISBN 9788522468959. Citado 3 vezes nas páginas 9, 10 e 11.

NI, Y. et al. A novel stock evaluation index based on public opinion analysis. In: **2018 International Conference on Identification, Information and Knowledge in the Internet of Things, IIKI 2018**. [S.l.]: Procedia Computer Science, 2019. v. 147, p. 581 – 587. Citado na página 21.

NONAKA, I.; KROGH, G. Tacit knowledge and knowledge conversion: controversy and advancement in organizational knowledge creation theory. **Organization Science**, v. 30, p. 635–652, 2009. Citado 2 vezes nas páginas 1 e 4.

NONAKA, I.; TAKEUCHI, H. **The Knowledge Creating Company: How Japanese Companies Create The Dynamics Of Innovation**. [S.l.: s.n.], 1995. v. 29. Citado 3 vezes nas páginas 1, 3 e 4.

O'LEARY, D.; STUDER, R. Knowledge management: an interdisciplinary approach. **IEEE Intelligent Systems**, v. 16, No. 1, 2001. Citado na página 1.

PENG, Z. Stocks analysis and prediction using big data analytics. In: . [S.l.: s.n.], 2019. p. 309–312. Citado 2 vezes nas páginas 23 e 25.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. **Information and Software Technology**, v. 64, p. 1–18, 2015. Citado na página 12.

PICASSO, A. et al. Technical analysis and sentiment embeddings for market trend prediction. **Expert Systems With Applications**, v. 135, p. 60–70, 2019. Citado na página 20.

PROBST, G.; RAUB, S.; ROMHARDT, K. **Gestão do Conhecimento: os elementos construtivos do sucesso**. Porto Alegre: Bookman, 2002. Citado 2 vezes nas páginas 3 e 4.

RODRIGUEZ-ELIAS, O. M. et al. A framework to analyze information systems as knowledge flow facilitators. **Information and Software Technology**, v. 50, p. 481–498, 2008. Citado na página 4.

ROMANOWSKI, A.; SKUZA, M. Towards predicting stock price moves with aid of sentiment analysis of twitter social network data and big data processing environment. In: _____. **Advances in Business ICT: New Ideas from Ongoing Research**. Cham: Springer International Publishing, 2017. p. 105–123. ISBN 978-3-319-47208-9. Disponível em: <https://doi.org/10.1007/978-3-319-47208-9_7>. Citado 3 vezes nas páginas 25, 27 e 29.

SEIF, M.; HAMED, E.; HEGAZY, A. Stock market real time recommender model using apache spark framework. In: _____. [S.l.: s.n.], 2018. p. 671–683. ISBN 978-3-319-74689-0. Citado na página 28.

SEZER, O.; OZBAYOGLU, M.; DOGDU, E. An artificial neural network-based stock trading system using technical analysis and big data framework. In: . [S.l.: s.n.], 2017. p. 223–226. Citado na página 32.

SHAH, D.; ISAH, H.; ZULKERNINE, F. Predicting the effects of news sentiments on the stock market. In: . [S.l.: s.n.], 2019. Citado 2 vezes nas páginas 21 e 27.

SHIH, P.-Y. et al. A rough petri nets model for stock trading signal detection. In: . [S.l.: s.n.], 2016. p. 135–139. Citado na página 31.

SISMANOGLU, G. et al. Deep learning based forecasting in stock market with big data analytics. In: . [S.l.: s.n.], 2019. p. 1–4. Citado na página 29.

SPENDER, J. C. Making knowledge the basis of a dynamic theory of the firm. **Strategic Management Journal**, v. 17, p. 45–62, 12 1996. Citado na página 3.

TANG, E.; FAN, Y. Performance comparison between five nosql databases. In: **2016 7th International Conference on Cloud Computing and Big Data (CCBD)**. Macau, China: IEEE, 2016. p. 105–109. Citado 2 vezes nas páginas 7 e 33.

THEOBALD, O. **Machine Learning for Absolute Beginners**. 1. ed. [S.l.: s.n.], 2017. Citado na página 8.

TSENG, C.-H. et al. Extraction, modeling, and predicting: a web driven approach for taiwan stock prediction. **Journal of the Chinese Institute of Engineers**, v. 41, p. 1–9, 11 2018. Citado na página 29.

VENKADACHALAM, R. et al. Back propagation neural network based big data analytics for a stock market challenge. **Communication in Statistics- Theory and Methods**, p. 1–21, 11 2018. Citado 2 vezes nas páginas 28 e 32.

WANG, Y.; WANG, Y. Using social media mining technology to assist in price prediction of stock market. In: **Proceedings of 2016 IEEE International Conference on Big Data Analysis, ICBDA 2016**. [S.l.: s.n.], 2016. Citado na página 23.

WEIHUA, X.; QIN, S. An analysis method of the business circumstance and stock price of listed company with social data. In: **2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)**. [S.l.: s.n.], 2018. p. 608–612. Citado na página 32.

WENG, B. et al. Predicting short-term stock prices using ensemble methods and online data sources. **Expert Systems with Applications**, v. 112, 06 2018. Citado na página 29.

YANG, H.; LIU, X.-Y.; WU, Q. A practical machine learning approach for dynamic stock recommendation. In: . [S.l.: s.n.], 2018. p. 1693–1697. Citado na página 28.

YANG, R. et al. Big data analytics for financial market volatility forecast based on support vector machine. **International Journal of Information Management**, v. 50, p. 452–462, 2020. ISSN 0268-4012. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0268401218313604>. Citado na página 27.

YOON, S.; SUGE, A.; TAKAHASHI, H. Do news articles have an impact on trading? - korean market studies with high frequency data. In: MINESHIMA, K. et al. (Ed.). **New Frontiers in Artificial Intelligence - JSAI-isAI Workshops, JURISIN, SKL, AI-Biz, LENLS, AAA, SCIDOCA, kNeXI, Revised Selected Papers**. [S.l.]: Springer Verlag, 2018. (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)), p. 129–139. ISBN 9783319937939. Citado na página 33.

ZAFAR, R. et al. Big data: The nosql and rdbms review. In: **International Conference on Information and Communication Technology (ICICTM)**. KUALA LUMPUR, Malaysia: [s.n.], 2016. p. 120–126. Citado 2 vezes nas páginas 5 e 7.

ZHU, X. **Semi-Supervised Learning Literature Survey**. [S.l.], 2005. Citado na página 9.